Topographic Maps of Semantic Space

William E. M. Lowe

Doctor of Philosophy

Institute for Adaptive and Neural Computation
Division of Informatics
University of Edinburgh

2000

Abstract

This work develops and tests the hypothesis that similarity of meaning derives from substitution regularities in the linguistic environment, represented topographically in the brain.

We develop a general mathematical theory of semantic space models and use this to motivate a set of new methods for high-dimensional space construction. We then develop a low-dimensional topographic map model of the lexicon from statistical considerations and apply both models to a range of semantic priming experiments.

We show that both high and low-dimensional models capture a wide range of previously un-modelled semantic relations. We also model the effects of association, semantic relatedness and their interaction, and offer a new theory of associative relations. We then demonstrate that, contrary to previous findings, the models also replicate graded and mediated priming effects. Mediated priming is of theoretical importance to memory models because its existence has been taken as evidence for spreading activation and against compound cue models. We show how a semantic space account representing only substitutability relations can account for mediated priming without making any specific architectural assumptions.

Acknowledgements

Thanks first to my supervisors, to Mark Ellison for inspiring me to think about topography and probability theory, to David Willshaw for giving me a home in his research group, and most of all to Richard Shillcock for trusting that I'd come up with something if left to my own devices. They all helped shape this thesis; some of them even came up with the money.

Thanks also to my various officemates, Gert Westermann, Sarah Gingell, and particularly to David Sterratt, good friend, best man, and unremittingly nice guy.

I have benefited enormously from the intellectual atmosphere in Edinburgh University's Centre for Cognitive Science, and in the Institute for Adaptive and Neural Computation. This work was partly completed while at the Center for Cognitive Studies at Tufts, for which I thank Daniel Dennett.

My final thanks go to Joanna Bryson. She knows what for.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*William E. M. Lowe*)

For JB, who probably won't read it.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "We shall say then that the meaning of a word is fully reflected in its contextual relations; in fact, we can go further, and say that, for present purposes, the meaning of a word is constituted by its contextual relations."
> D. Cruse (1986) *Lexical Semantics*

This is a thesis about going further. The following chapters formulate, operationalise and test the following hypothesis:

> Similarity of meaning derives from substitution regularities in the linguistic environment, represented topographically in the brain.

Wittgenstein (1958) argued that words are not similar in meaning because each word is associated with an abstract object, its *meaning*, that can be compared with other meanings and found to alike. Rather words are similar in meaning when they are used in similar ways. But as it stands, this formulation of meaning as use puts no constraints on semantic theory; everyone agrees that whatever it is that makes words similar in meaning will affect the way they are used. Indeed it seems that Wittgenstein has it backwards. Surely words are used in similar ways *because* they mean similar things — 'doctor' is semantically related to 'nurse' because doctors and nurses work together in hospitals, not because the words themselves tend to share similar sentential contexts. But although this perspective is not intuitive it is extremely powerful: contextual similarity between words has already provided explanations of a large number of psychological variables. In Chapter 5 we add to that collection.

If Wittgenstein's approach is correct then the psychological question is, how best to represent contextual similarity in a computational model? Semantic space models represent similarity of use, and by hypothesis similarity of meaning, by angular structure

or distance in a high-dimensional space defined by word co-occurrence counts. Chapter 3 places them in context with alternative modelling frameworks for understanding semantic memory and priming. Chapter 4 presents a general detailed theory of semantic space motivated by the idea of a generalised replacement test and introduces new statistically-motivated methods for constructing semantic space models. We also analyse current semantic models and show why simple co-occurrence counts are not desirable for semantic space models.

Current models assume that semantic space is of very high dimensionality. It is then important to understand how such a space could be represented within the constraints of neural tissue. Chapter 2 introduces the Generative Topographic Mapping, a statistical latent variable model that creates neural maps. Topographic maps model high-dimensional data as generated by a latent or unobserved space of low dimensionality. In Chapter 5 we show how angular structure on a map surface can recreate many of the psychological effects previously only dealt with in high dimensions. Success with topographic maps is taken as evidence that the intrinsic dimensionality of semantic space is in fact low, and that maps therefore constitute a plausible neural implementation of semantic space.

Chapter 5 applies high and low-dimensional semantic space models to five semantic priming studies. In Experiments 1 to 4 we investigate the nature of associative priming and its interaction with semantic priming. In the next four studies we show that both high and low-dimensional models capture priming due to a wide range of previously unexplored semantic relations, and replicate interactions between semantic relatedness and association. Previous accounts of associative relations have assumed a conditional probability theory – two words are associated if the occurrence of one is made more probable by the prior occurrence of the other. We present a novel alternative theory of associative priming that explains how a semantic space that only reflects substitutability in context can capture both associative and semantic priming despite having no mechanism for representing conditional probability. We then address graded and mediated priming effects. In Chapter 3 we consider why mediated priming effects have been argued to be critical in deciding between classical spreading activation and compound cue models of semantic memory. Also, previous attempts to model mediated priming in semantic space have failed, leading Livesay and Burgess (1998) to argue that the effect cannot be explained by non-mediated means, and therefore not by a semantic

space. We show how mediated priming is a special case of graded priming effects due to weak but direct relatedness, and model it successfully in a high-dimensional space.

The contributions of this research are: a general mathematical theory of semantic space models, a synthetic review of topographic map models, a set of new methods for space construction, the development and testing of a topographic map model of the lexicon using real psychological data, a demonstration that a wide range of previously un-modelled semantic relations are represented in semantic space, a new theory of associative priming that explains how conditional probability affects substitutability estimates, a demonstration that graded and mediated priming can be captured in a semantic space model.

Reading the thesis

This thesis was written with a interdisciplinary audience in mind, so it is perhaps inevitable that not everything will be of equal interest to all readers. The following suggestions might make for more profitable reading.

Readers interested in seeing how effectively semantic space models can be applied to psychological data may find it useful to skip over chapter 2, read chapter 3, skim chapter 4 and then examine the results in chapter 5. Readers of a more theoretical bent who are interested in the theory and assumptions underlying semantic space modelling might skim chapter 2, and concentrate on chapters 3 and 4 where most of the psycholinguistically relevant theory is developed. These readers might also be interested to see how notions such as latent variable are applicable across disciplines; if so they should also read chapter 2. Chapter 2 can also stand alone as synthetic review of the topographic mapping literature in applied statistics and neural network research. Readers interested purely in evaluating the models developed here against human data and alternative theories can skip chapter 2, skim chapters 3 and 4, and immerse themselves in the priming data of chapter 5.

Chapter 2

Latent Variable Models and Topographic Maps

The Generative Topographic Mapping (GTM Bishop et al., 1998; Svensén, 1998) is a statistical model for generating vectors of real valued observations on the basis of unobservable or latent variables. The GTM also creates topographic maps; high-dimensional data points are represented by locations on a low dimensional manifold such that nearby points are mapped to neighbouring locations[1]. This chapter develops the GTM as a member of the class of statistical latent variable models similar to Factor Analysis. We show how topographic representation arises naturally from probabilistic inference on these models. We then consider some popular and widely studied alternative map models and show how each model is an approximation to or special case of a generalisation of the GTM that uses Gaussian Processes. Neural interpretations of model parameters and structure are provided throughout the chapter.

Integrating Neural and Statistical Approaches to Topographic Maps

Combining theory developed in the cortical map literature with statistical theory is difficult in part because most neural network models of topographic maps specify a mapping *from* the data space *onto* the lower-dimensional map surface, whereas latent variable models define a mapping in the reverse direction.

[1]Manifold here denotes a smooth continuous mapping from an open interval in $\mathcal{R}^L$ to $\mathcal{R}^D$. Of particular relvance here is the case where $L < D$. When the mapping is also linear, the manifold is a closed region in an L-dimensional subspace of $\mathcal{R}^D$. When the mapping nonlinear, the manifold can be intuitively understood as a rubber sheet that may twist and stretch across the dimensions of $\mathcal{R}^D$.

In specifying a mapping into a lower dimensional space it is natural to expect inter-point distances, or at least distance ranks, to be approximately preserved. This leads to a class of solutions, e.g. Multidimensional Scaling (MDS; Torgerson, 1952; Shepherd, 1962) and Sammon Mapping (Sammon Jr., 1969), that minimise an objective function based on inter-point distance correlations between high-dimensional points and their low dimensional projections (Goodhill and Sejnowski, 1997). Direct minimisation is a difficult non-linear optimisation problem, but more problematically, the resulting projection is only defined for the original set of data points; to project more additional data the optimisation process must be repeated.

From a statistical point of view MDS is also unsatisfactory because it does not define a distribution in the lower-dimensional space; it is therefore not possible to express uncertainty about the low-dimensional position of a data point. From a biological point of view MDS can be seen as an abstract neural model; for example a Classical MDS solution to the problem of mapping into two dimensions is to take the first two principal components of the data, which can be implemented as a Hebbian learning scheme (Oja, 1989; Hertz et al., 1991). However, it is not obvious that MDS helps understand how or what brains are mapping, save that they should be doing it topographically. MDS is also unsatisfactory from a statistical or psychological viewpoint. Non-metric MDS is based on a pure optimisation process and has no associated statistical model. In the absence of a probabilistic expression for observation noise generalisation to new data is not defined; each new datum requires the optimisation process to be repeated. There is also little control over the flexibility of the final mapping.

If the problem of topographic mapping is considered as one of mapping low-dimensional points into a higher dimensional space, it is no longer natural to require that inter-point distances or ranks are preserved. Then it is possible to assign distributions to the latent variable[2], to the parameters of the mapping into high dimension and therefore also to the resulting position in high-dimensional space. The presence of a latent distribution makes it straightforward to invert the mapping using Bayes theorem, project any new data point onto the map surface and express the level of uncertainty associated with the projected position. This reformulation of the topographic mapping problem is explored in detail below.

[2]Distributional assumptions are most illuminating when the mapping is in this direction because they directly specify the nature of the data generation mechanism, whereas placing a distribution directly on the high dimensional points makes less sense; if we knew how they were distributed it would not be necessary to do exploratory data analysis at all.

The development of latent variable formulations of neural network models coincides with a 'generative turn' in neural networks research (e.g. Hinton and Ghahramani, 1997; Neal, 1996; Bishop, 1995; MacKay, 1991). The generative turn is a methodological approach to neural representation in which the brain's task is to learn a generative model of the structure of its environment; perception is then equivalent to inverting the generative model (Knill and Richards, 1996).

It is useful to situate the GTM with respect to more familiar statistical models and their recent extensions. The following section presents a sequence of latent variable models of increasing complexity that are important for understanding the GTM. The section begins with classical factor analysis (FA), develops probabilistic principal component analysis (PPCA) and ends by presenting the GTM as a non-linear generalisation of PPCA.

2.1 Latent Variable Models

Latent variable models assume that, although each data point $\mathbf{t} = [t^1 \dots t^D]^\mathsf{T}$ consists of D measurements, they are not all necessary to explain the observed structure of the data, and that the *intrinsic dimensionality* is actually lower. Typically latent variable models reflect the most variant directions of the data set[3], whether or not they coincide with the dimensions in which the data were originally measured (notice that latent structure in Figure 2.1 is not aligned with the axes). Figure 2.1 shows how variance can be used as a guide to finding underlying structure in some very simple data. Models differ with respect to what assumptions they make about the number and structure of their latent variables and the type of noise affecting measurement. We begin by considering Factor Analysis, perhaps the simplest latent variable model of the relation between the intrinsic dimensionality of the data and what is observed.

2.1.1 Factor Analysis

In factor analysis a D-dimensional data point $\mathbf{t}$ is assumed to be the result of choosing a point from an L-dimensional Normally distributed latent variable $\mathbf{x} = [x^1 \dots x^L]^\mathsf{T}$ where

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}), \tag{2.1}$$

[3]although Independent Component Analysis (Bell and Sejnowski, 1995) is a recent exception.

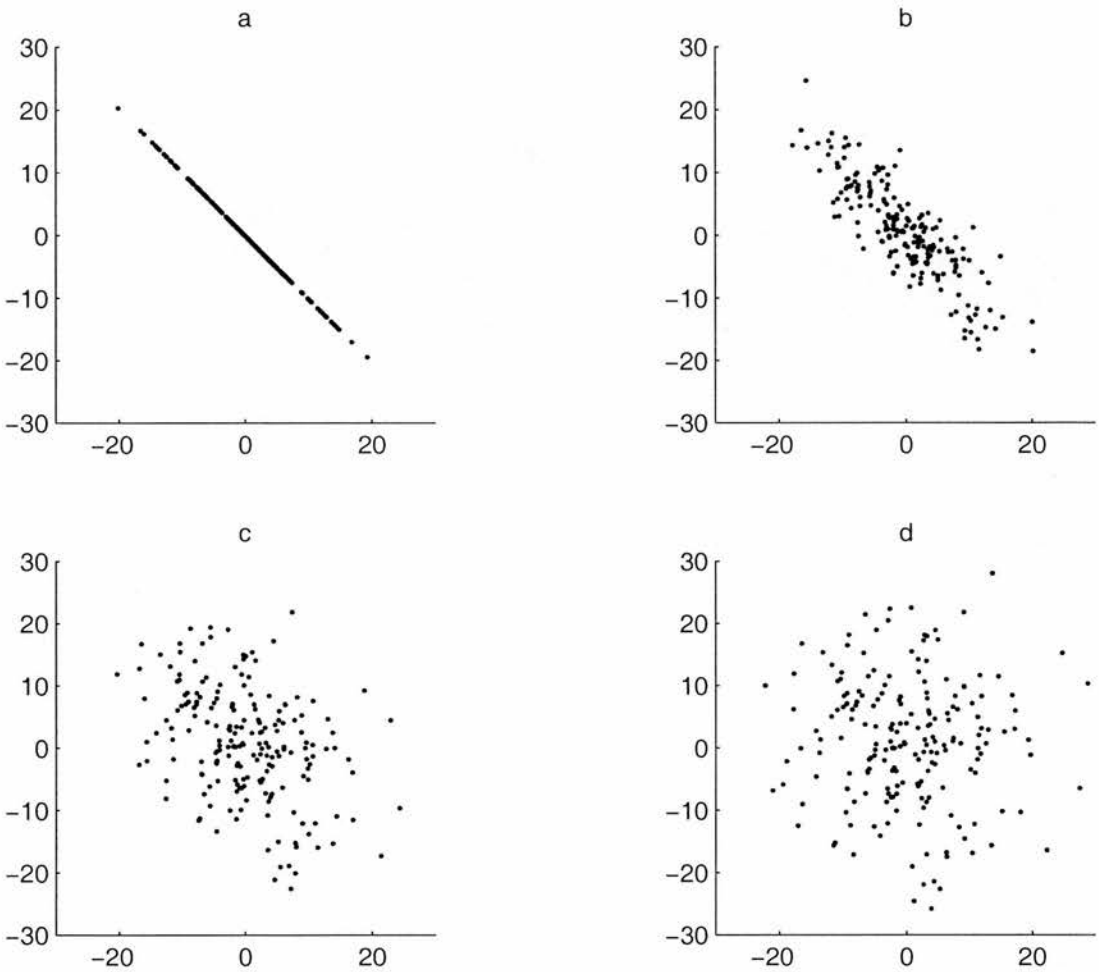

Figure 2.1: Two hundred data points sampled from a one-dimensional Normal distribution $\mathcal{N}(\mu, \sigma^2)$ over the line $y = -x$, where μ=0 and σ=10. In a) the data is confined to a one-dimensional subspace of the data space and thus has intrinsic dimensionality 1; x and y values are perfectly negatively correlated (r=-1). In b) the data is perturbed by additive zero mean Gaussian noise with variance ς=3, perpendicular to the $y = -x$ plane. Strictly the intrinsic dimensionality of the data is now 2, but because $\varsigma \ll \sigma$, x and y are highly correlated (r=-.85) and a model that ignores the direction of smallest variance constitutes a good approximation to the data structure. c) denotes a similar situation where ς=7; r=-.39, so the intrinsic dimensionality is more obviously 2. In d) ς=σ and r=-.05, so the intrinsic dimensionality is clearly 2 and a model that ignores the direction of smaller variance is inadequate for representing the data structure.

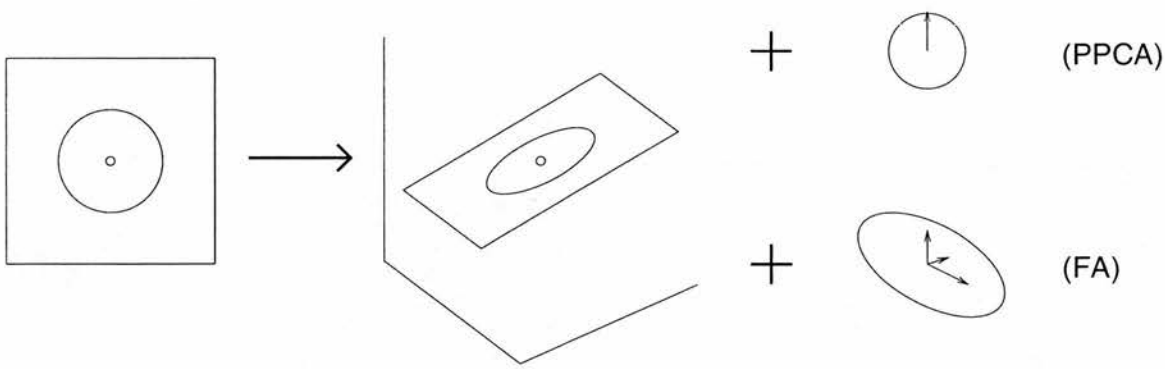

Figure 2.2: A cartoon of Probabilistic Principal Component Analysis and Factor Analysis as latent variable models. Both models map a zero mean unit variance random variable (left) into a potentially higher dimensional data space (centre), producing a distribution over some subspace. The PPCA model then adds spherical noise (top right ellipse) to every point in the subspace to give a probabilistic model for the data. In FA the noise model is axis-aligned and may have different variances (bottom right ellipse). A Monte Carlo approximation to PPCA or FA chooses a finite number of points in latent space, maps them into the data space and applies the appropriate noise model to each point individually to give a mixture model (see later).

(Figure 2.2, left. See Appendix B for a guide to notation), mapping it linearly into the data space (Figure 2.2, centre) and adding Normally distributed measurement noise (Figure 2.2, bottom right). For simplicity of exposition we can take the mean of data to be zero since this doesn't affect the variance structure. The generative model is then

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon} \tag{2.2}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \tag{2.3}$$

$\mathbf{W}$ is a real-valued $D \times L$ matrix, called the factor loadings. $\boldsymbol{\epsilon}$ represents the measurement errors from D measurements,

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2 \ldots \sigma_D^2)$$

where $\text{diag}(a_1 \ldots a_D)$ represents a matrix with a_1 to a_D on the main diagonal and zeros everywhere else.

In psychological applications $\mathbf{t}$ may be a set of D psychometric measurements or test results for an individual. Then the latent variable represents the underlying L psychological traits $\mathbf{x}$ that give rise to the observed correlations between test results (Everitt, 1984). An explanatory theory will assume fewer independent traits than tests so $L < D$. Factor analysis assumes that linear combinations of traits fully account for correlations between test results so that any residuals, $\sigma_1^2 \ldots \sigma_D^2$ are due to noise inherent in each test. The noise model for Factor Analysis is represented by the ellipse in the bottom half of Figure 2.2. The noise is axis-aligned representing the fact that the dimensions of each data point are conditionally independent given knowledge of the value of the latent variable,

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{W}\mathbf{x}, \boldsymbol{\Sigma}). \tag{2.4}$$

However, the latent variable is by definition unobserved so its value is uncertain. The standard Bayesian, and in this case also Classical, approach to uncertainty about the value of a variable is to integrate over all of its possible values, a process known as marginalisation. The full Factor Analysis model of a single datum is then

$$p(\mathbf{t} \mid \mathbf{W}, \boldsymbol{\Sigma}) = \int p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \boldsymbol{\Sigma})\, p(\mathbf{x})\, d\mathbf{x} \tag{2.5}$$

It is straightforward to calculate the probability of a data set $\mathbf{T}$ consisting of N independent data points under the generative model,

$$p(\mathbf{T} \mid \mathbf{W}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} p(\mathbf{t}_i \mid \mathbf{W}, \boldsymbol{\Sigma}) \tag{2.6}$$

The right hand side of Equation 2.6 can be treated as a function of $\mathbf{W}$ and $\mathbf{\Sigma}$, and maximised to obtain Maximum Likelihood values for the model parameters. Parameters for all of the latent variable models discussed in this chapter may be set using Expectation Maximisation (EM) algorithms (Dempster et al., 1977). Although the focus of this chapter is on the assumptions each model makes about the latent structure of data, rather than on parameter fitting methods, it is worth emphasising that although the EM algorithm has a strong statistical motivation, it is in fact also a very general form of unsupervised learning procedure; many unsupervised neural network training algorithms are special cases.

Monte Carlo Sampling

The random variables in Equations 2.4 and 2.1 are both Normally distributed so the integral in Equation 2.5 can be solved analytically,

$$p(\mathbf{t} \mid \mathbf{W}, \mathbf{\Sigma}) \;=\; \mathcal{N}(\mathbf{W}\mathbf{x}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \mathbf{\Sigma}). \tag{2.7}$$

However in preparation for later, it is useful to see how the integral in Equation 2.5 would be treated numerically. This would be necessary, for example, if the model did not assume that $\mathbf{x}$ was Gaussian distributed, or the model was to be implemented in a network of discrete neurons. We consider neural interpretations of latent variable models in detail below.

A Monte Carlo estimate of Equation 2.5 starts by sampling M points $\mathbf{x}_i$ from the latent variable. The points may be chosen randomly according to the distribution of $p(\mathbf{x})$. For Factor Analysis points are chosen from a standard Normal distribution, although for more complex latent spaces a more involved sampling scheme may be necessary (Neal, 1993). Each point is then mapped by the loading matrix into the data space and treated as the mean of a Normal distribution with covariance matrix $\mathbf{\Sigma}$. The numerical approximation is then

$$p(\mathbf{t} \mid \mathbf{W}, \mathbf{\Sigma}) \;\approx\; \frac{1}{M} \sum_{i=1}^{M} \mathcal{N}(\mathbf{W}\mathbf{x}_i, \mathbf{\Sigma}). \tag{2.8}$$

As $M \to \infty$ and the distance between samples tends to zero, the approximation becomes exact.

Posterior inference is slightly more involved using a Monte Carlo approximation: the true expected position for a datum in latent space may not be among the chosen

sample points $\mathbf{x}_1 \ldots \mathbf{x}_M$. Further, substituting Equation 2.8 into Bayes theorem only gives posterior probabilities for $\mathbf{x}_1 \ldots \mathbf{x}_M$. One solution is to choose the sample point most likely to have generated the datum. This may not be a good approximation if M is small or unevenly spaced. Alternatively an approximation to $\mathbf{Bt}$ can be constructed by taking a linear combination of latent sample points, weighted by their posterior distribution,

$$\langle \mathbf{x} \mid \mathbf{t} \rangle = \sum_{i=1}^{M} \mathbf{x}_i \, p(\mathbf{x} \mid \mathbf{t}, \boldsymbol{\Sigma})$$

This approximation *can* take values not equal to the latent sample points and will also converge to the correct value as M increases and the distance between sample points decreases. The variance of the estimate is computed similarly.

Computing $\langle \mathbf{x} \mid \mathbf{t} \rangle$ is also a good example of the advantages of coarse coding: the set of points in latent space that can be distinguished by a linear combination of sample points is typically much larger than the set of sample points. This property is a consequence of the broad 'tuning' of the noise model controlled by $\boldsymbol{\Sigma}$ (McClelland and Rumelhart, 1988).

2.1.2 Probabilistic Principal Component Analysis

In Factor Analysis, data are modelled as the result of a single Normally distributed latent variable, centred on the data mean and spanning a subspace of the data space, perturbed by axis-aligned Normal measurement noise. Probabilistic Principal Component Analysis (PPCA; Tipping and Bishop, 1997; Roweis, 1998) is a similar model based on classical principal component analysis (Joliffe, 1986; Jackson, 1991). PPCA is also closely related to the Latent Semantic Indexing model of Landauer and colleagues, and is described in more detail there.

Both Factor Analysis and PPCA involve a linear mapping from a latent variable into the data space. They differ only in the assumptions they make about noise. The following sections extend the class of latent variable models by allowing other choices of latent variable, and more flexible mappings. In preparation for later use it is useful to rewrite Equation 2.2 functionally,

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}) = \mathcal{N}(\mathbf{y}(\mathbf{x}; \mathbf{W}), \boldsymbol{\Sigma}) \tag{2.9}$$

$$\mathbf{y}(\mathbf{x}; \mathbf{W}) = \mathbf{W}\mathbf{x}. \tag{2.10}$$

This notation emphasises the fact that other choices of functional mapping are possible by separating the projection of a latent point into the data space from its associated noise distribution.

2.2 Posterior Inference and Maps

It is often useful to infer the value of the latent variable for different data points. In a psychological context this corresponds to inferring the psychological traits of an individual on the basis of their test results and a fitted generative model. If $L < 3$, points in the latent space can be plotted (see Figure 2.3). The inverse mapping from data to latent space is obtained by Bayes theorem using Equations 2.1 and 2.4.

$$p(\mathbf{x} \mid \mathbf{t}, \mathbf{W}, \boldsymbol{\Sigma}) = \frac{p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \boldsymbol{\Sigma})\, p(\mathbf{x})}{p(\mathbf{t} \mid \mathbf{W}, \boldsymbol{\Sigma})} \tag{2.11}$$

where the denominator is the left hand side of Equation 2.5. Equation 2.11 provides a posterior distribution of values for the latent variable given any particular data point. In the Factor Analysis model the posterior distribution has a closed form (Roweis and Ghahramani, 1999).

Plotting the expected posterior values of each data point in the latent space then creates a reduced-dimension *map* of the data set. This interpretation of the posterior distribution over a low dimensional latent space as a map provides the crucial connection between generative statistical models and neural network topographic map models. The map that results from posterior inference is an explicit representation of the underlying variance structure in the data.

2.3 Introducing Non-linearity

Neither Factor Analysis nor PPCA can represent essentially non-linear data. Figure 2.3a shows a curved generating distribution with intrinsic dimensionality 2, spread out in three dimensions. Figure 2.3b shows 150 data points sampled from the generating distribution with additive spherical Gaussian noise. The two circled points are at opposite ends of the generating distribution. When a PPCA model with L=2 is fitted to the data it follows the directions of maximum variance; in this case they are not a good guide to the latent structure (the square in Figure 2.3c is an arbitrary plane for orienting the projection visually). Ellipses are lines of constant probability

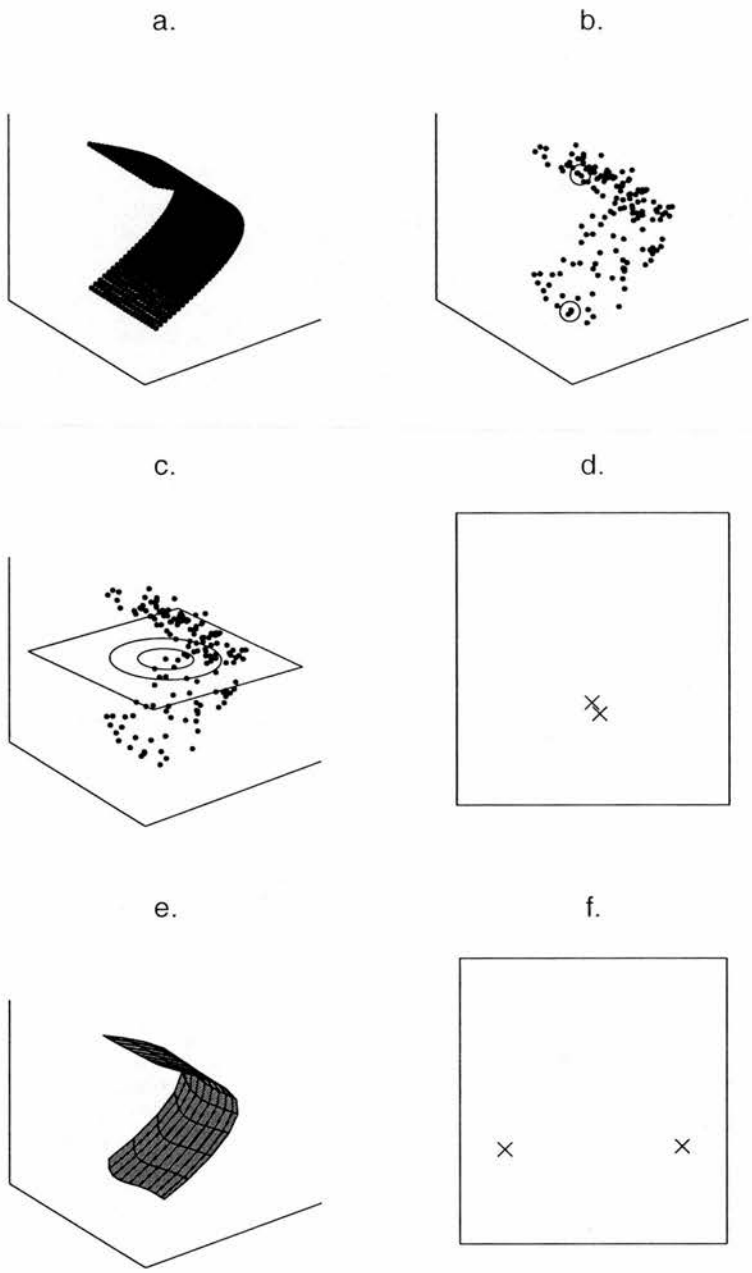

Figure 2.3: Data of intrinsic dimensionality 2. a) Underlying planar structure in the data. b) 150 samples corrupted by Gaussian noise. Two distant samples are circled. c) the fitted PPCA plane; ellipses are lines of constant probability at one and two standard deviations from the mean. The square shows the orientation of the latent subspace. d) Distant samples are confounded in the posterior projection. e) A fitted GTM model, displayed in the data space. f) Posterior means for distant samples in the latent space.

at one and two standard deviation intervals around the data mean. Figure 2.3d shows the posterior mean positions for the two data points circled in b). Their positions are confounded in the posterior projection because they differ in a direction orthogonal to that spanned by the PPCA subspace. If the data is clustered, the problem can be addressed using a latent variable model that defines multiple local models of the data. The following sections describe alternative latent variable models designed to capture non-linear underlying structure.

2.4 Mixture Models

Perhaps the simplest choice of local latent variable model is a mixture of Gaussians with shared spherical covariances. This model assumes that data is generated by exactly one of M Normal distributions positioned in the data space with means $\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_M$ and covariance matrices $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Each data point is associated with a multinomially distributed latent variable $\boldsymbol{\omega} \sim \mathcal{M}(\boldsymbol{\pi}, 1)$ that identifies the generating distribution. $\omega_i = 1$ and $\omega_j = 0, j \neq i$ when distribution i generated the datum. $\boldsymbol{\omega}$ is itself controlled by a vector of multinomial prior probabilities $\boldsymbol{\pi} = [\pi_1 \ldots \pi_M]$. In the following $p(\omega_i = 1)$ is abbreviated as $p(\omega_i)$ and is equal to π_i.

If the generating distribution i is known then the probability of a data point is given by its probability under the generating distribution:

$$p(\mathbf{t} \mid \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma}, \omega_i) = \sum_{j=1}^{M} \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})\, \omega_j \tag{2.12}$$

$$= \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \tag{2.13}$$

However the identity of the generating distribution is unknown so the probability of a data point under the mixture model is obtained by marginalising over the prior on $\boldsymbol{\omega}$

$$p(\mathbf{t} \mid \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) = \sum_{i=1}^{M} p(\mathbf{t} \mid \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma}, \omega_i)\, p(\omega_i) \tag{2.14}$$

$$= \sum_{i=1}^{M} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\, \pi_i \tag{2.15}$$

There is no constraint on the positions of the means in data space, so the mixture model can capture cluster structure in the data.

To infer which distribution is most likely to have generated a data point, the model is inverted using Bayes theorem to give a posterior distribution over values of $\boldsymbol{\omega}$,

$$\begin{aligned} p(\omega_i \mid \mathbf{t}, \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) &= \frac{p(\mathbf{t} \mid \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma}, \omega_i)\, p(\omega_i)}{p(\mathbf{t} \mid \{\boldsymbol{\mu}\}, \boldsymbol{\Sigma})} \\ &= \frac{\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})\, \pi_i}{\sum_{j=1}^{M} \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})\, \pi_j} \end{aligned} \tag{2.16}$$

Equation 2.16 shows for each component distribution i the probability that it generated $\mathbf{t}$. It is easy to see that as $\sigma^2 \to 0$, Equation 2.16 reduces to a nearest neighbour rule: the posterior most probable mean for $\mathbf{t}$ is the one nearest to it in the data space. This observation connects the mixture model framework to vector quantisation, and thus to competitive learning algorithms for neural networks (Hertz et al., 1991; Ritter et al., 1991).

2.4.1 Interpreting mixture models

Equation 2.14 is expressed as a sum over distributions because the mixture model represents the probabilistic equivalent of an exclusive-or: either i generated the datum *or* j generated it *or* k etc. but it is not possible that *both* i and j generated it because these possibilities are mutually exclusive. It is only because these events are mutually exclusive that the marginalisation expressed in Equation 2.14 makes probabilistic sense. Consequently the spread of probability over different values of $\boldsymbol{\omega}$ in the posterior distribution must be interpreted as our uncertainty about the identity of the generating distribution, rather than an inference about the degrees to which each distribution took part in the generation process.

One important consequence of the mixture model's exclusive-or semantics is that it does not make sense to map a datum to a position in latent space that is specified by a weighted average of means, where the weights are posterior probabilities of having generated that point; in other words, the latent variable is discrete, so averaging it makes no sense. However, section 2.1.1 presents a Monte Carlo approximation scheme that is algebraically equivalent to a mixture model with $\pi_i = 1/M$ (compare Equations 2.14 with 2.8), and recommends averaging over posterior probabilities to obtain the expected position of a data point in latent space. How should these contradictory-seeming recommendations be reconciled[4]?

[4]This section is motivated by an objection from Geoffrey Hinton (pers. comm.)

The essential difference between Factor Analysis and a mixture of Gaussians is in the nature of the latent variable: in Factor Analysis $\mathbf{x}$ is continuous and in the mixture of Gaussians $\boldsymbol{\omega}$ is discrete. A Monte Carlo approximation to Factor Analysis takes the form of a mixture model simply because it is a finite element approximation to an integral – the underlying space is still continuous. In a mixture of Gaussians however there is no approximation involved and therefore no underlying space. The status of the $\mathbf{x}_i$ is also quite different to the mixture model's $\boldsymbol{\omega}$. The identity of the $\mathbf{x}_i$ can be changed at any time and M can be increased or decreased according to computational convenience. In contrast, M is fixed in the mixture model[5]. These differences mean that any linear combination of $\mathbf{x}_i$ is guaranteed to be a possible position in latent space, which justifies computing $\langle \mathbf{x} \mid \mathbf{t} \rangle$, whereas no non-trivial linear combinations of mixture model latent variables will be possible values of $\boldsymbol{\omega}$. This distinction is relevant for understanding the use of $\langle \mathbf{x} \mid \mathbf{t} \rangle$ in the GTM.

2.5 Generative Topographic Mapping

Returning to Figure 2.3, the generating curve could be approximated by a mixture of Gaussians, particularly if their covariance matrices were altered to be able to reflect planar structure. Each mixture element could then deal with a different locally linear section of the data structure. In the example two Gaussians would provide a reasonable approximation. However, a mixture representation does not reflect the intrinsic dimensionality of the generating distribution. For example, the more non-linear the data structure is, the more mixture elements will be necessary to model it, even if it is still a smooth manifold of low dimension. Also when multiple distributions are used to model the data it can no longer be globally visualised. Each data point can be assigned to the mixture element with the highest responsibility, but it cannot then be projected into a lower-dimensional latent space, since none exists for mixture models. The GTM is an attempt to remedy these problems by fitting a low dimensional but non-linear manifold to the data. A side effect of fitting a low dimensional manifold is that all data points can be projected into the same latent variable space to form a map.

The GTM fits an L dimensional manifold to the data – for visualisation purposes

[5] 'Constructivist' neural networks (Frean, 1990; Fritzke, 1994) are apparently an exception to this claim, since they increase M according to an error criterion. However, the probabilistic basis of these models is unclear.

typically L=2. The latent space is given by a uniform distribution over an arbitrary open interval in $\mathcal{R}^L$,

$$\mathbf{x} \sim \mathcal{U}(-1, 1) \tag{2.17}$$

To create the necessary non-linearity the linear mapping of Equation 2.10 is replaced by a generalised linear model

$$\mathbf{y}(\mathbf{x}; \mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \tag{2.18}$$

where $\boldsymbol{\phi}(\mathbf{x}) = [f_1(\mathbf{x}) \ldots f_M(\mathbf{x})]^{\mathsf{T}}$ is a vector of basis function outputs. Equation 2.18 describes a radial basis function network (RBF; Moody and Darken, 1989) and as such has universal approximation capability: with a suitable number of basis functions any mapping into the data space can be represented. The Bayesian treatment of neural networks requires that all weight parameters have a prior distribution. For the GTM this is expressed as a Gaussian prior favouring small weight values parameterised by inverse variance α:

$$p(\mathbf{W} \mid \alpha) = \mathcal{N}(0, \alpha^{-1}\mathbf{I}).$$

α is effectively a weight-decay regulariser that ensures that the mapping into the data space is smooth. Larger values of α force smaller weights which give smoother mappings into the data space.

Like a standard RBF and PPCA, the GTM has a spherical noise model. Following the original papers it is parameterised by an inverse variance parameter β. This may either be set by Maximum Likelihood, or used as an annealing parameter (see below).

If the value of $\mathbf{x}$ for datum $\mathbf{t}$ is known, then

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{y}(\mathbf{x}; \mathbf{W}), \boldsymbol{\Sigma}) \tag{2.19}$$

However, the value of $\mathbf{x}$ is unobserved, so the probability of $\mathbf{t}$ under the GTM is obtained by marginalising over the identity of the generating distribution (c.f. Equation 2.14),

$$p(\mathbf{t} \mid \mathbf{W}, \beta) = \int p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \beta)\, p(\mathbf{x})\, d\mathbf{x} \tag{2.20}$$

This integral is intractable, so a numerical approximation is obtained by making a Monte Carlo approximation to $p(\mathbf{x})$: M latent variable samples $\{\mathbf{x}\}$ spaced regularly

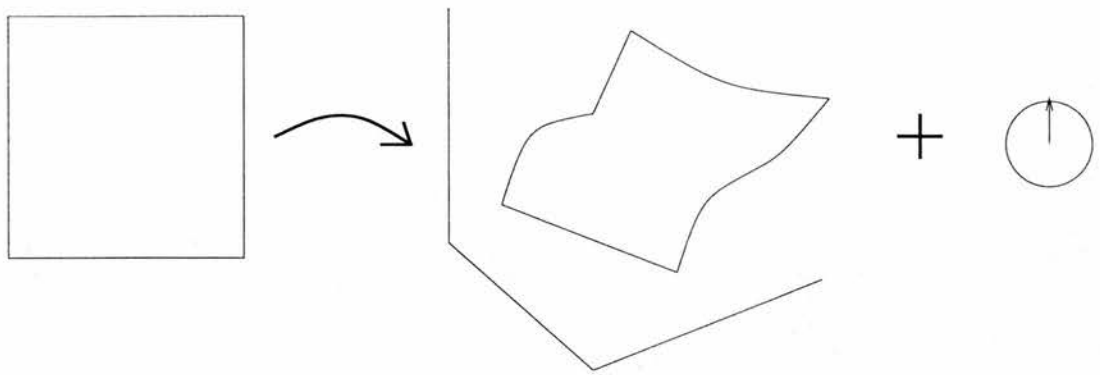

Figure 2.4: A cartoon of the Generative Topographic Mapping (Compare to Figure 2.2). A uniformly distributed random variable (left) is mapped non-linearly into the data space (centre) to create a manifold. Spherical noise (right) is then added to each point on the latent space to give a probabilistic model for the data. The delta prior (Equation 2.21) implements a Monte Carlo approximation to the full model; a finite number of points are chosen in latent space. Each point is mapped into the data space and assigned a spherical noise distribution, giving a mixture model.

in a grid define an approximation to the true latent space distribution. Bishop et al., 1998 describe the approximation as a sum of delta functions,

$$p(\mathbf{x}) \approx \frac{1}{M}\sum_{i=1}^{M}\delta(\mathbf{x}-\mathbf{x}_i). \tag{2.21}$$

However, this is very a misleading expression. As discussed above, this prior will not allow any positive probability in the posterior distribution to points that are not in the original latent sample. Thus Equation 2.21 defines a latent space that has the same problems of interpretation as the mixture model. The latent points should instead be interpreted as a sample from the true uniform prior distribution.

The marginal distribution of Equation 2.20 can be approximated by an M element mixture model

$$p(\mathbf{t} \mid \mathbf{W}, \beta) \approx \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(\mathbf{y}(\mathbf{x}_i;\mathbf{W}), \mathbf{\Sigma}) \tag{2.22}$$

(c.f. Equation 2.8). Each sample from the latent space has a weight of 1/M because the true latent space is a uniform distribution.

The GTM effectively defines a constrained mixture of Gaussians. It is constrained because each mean $\mathbf{y}(\mathbf{x}; \mathbf{W})$ is determined by the mapping from the latent space into the data space rather than being specified directly in the data space. The basis functions ϕ are typically smooth unimodal functions with infinite support, e.g. Gaussians, so although the latent variable is only sampled at discrete points, the model defines a smooth and continuous manifold in the data space.

2.5.1 Inversion

The model is inverted using Bayes theorem in the form

$$p(\mathbf{x}_i \mid \mathbf{t}, \mathbf{W}, \beta) = \frac{\mathcal{N}(\mathbf{y}(\mathbf{x}_i; \mathbf{W}), \mathbf{\Sigma})}{\sum_{j=1}^{M} \mathcal{N}(\mathbf{y}(\mathbf{x}_j; \mathbf{W}), \mathbf{\Sigma})}. \tag{2.23}$$

where the mixture weights $1/M$ have cancelled. To represent all the data points at once in the latent space, the point in latent space with the highest posterior probability for each data point can be shown: data point $\mathbf{t}_j$ is then represented by the latent point $\mathbf{x}_i$ when

$$i = \text{argmax}(a)\, p(\mathbf{x}_a \mid \mathbf{t}_j, \mathbf{W}, \beta).$$

Alternatively the posterior mean can be used,

$$\langle \mathbf{x} \mid \mathbf{t}_j \rangle = \sum_{i=1}^{M} \mathbf{x}_i\, p(\mathbf{x}_i \mid \mathbf{t}_j, \mathbf{W}, \beta).$$

Figure 2.3 shows the manifold from a fitted GTM model for the curve data. The mean posterior positions of the two circled data points are shown in f. They are well-separated because the model can accurately capture the planar structure of the generating distribution.

That the position of the posterior mean need not coincide with any of the $\{\mathbf{x}\}$ suggests a useful neural interpretation for the posterior distribution as a population code (Zemel et al., 1998). In a population code, the responses of a large number of widely tuned neurons are combined to yield a response that is more accurate than any single neuron's output. Population coding analyses are most often found for motor systems (Georgopoulos et al., 1988), but one widely discussed sensory application is in explaining vernier hyperacuity (Weiss et al., 1995; Wilson, 1986): Depending on the task, when subjects are presented with two vertical lines placed one upon the other

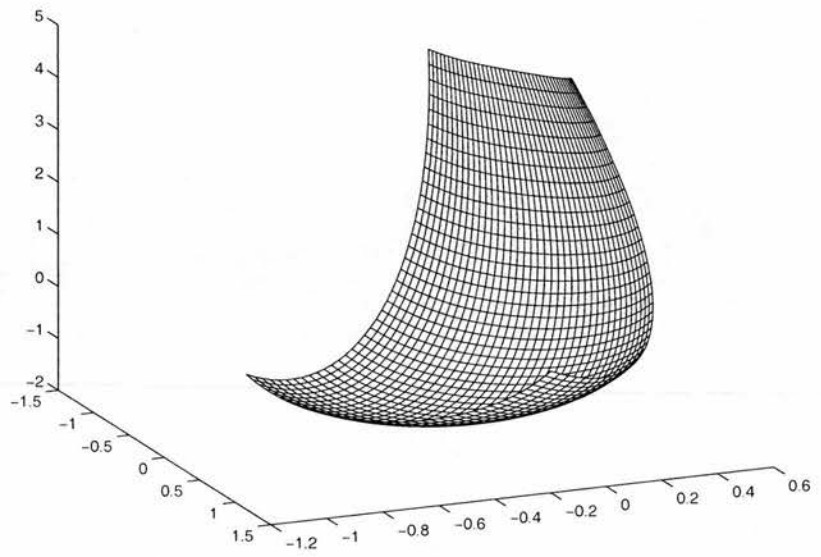

Figure 2.5: A sample manifold from inflexible GTM. 1600 latent points in two dimensions are passed through 16 Gaussian basis functions with means two standard deviations apart, and mapped randomly into data space. Neighbouring means in latent space are connected by vertices.

with a small horizontal offset at their join, they reliably resolve differences in offset that are one fifth the size of the spacing between photo-receptors in the eye (Edelman and Weiss, 1995). Accuracy beyond the limits of the available sensors is achieved by averaging many broadly tuned responses.

2.5.2 Topographic Mapping

The number and spacing of basis functions controls the flexibility of the topographic mapping (see Figures 2.5, 2.6 and 2.7). The mapping is naturally topographic because nearby points in the input space are necessarily mapped to nearby points in the basis, and the next layer of adjustable weights can only perform linear transformations which are guaranteed to maintain neighbourhood relations.

A small number of heavily overlapping basis functions forces the manifold in data space to be close to the planar mapping defined by the first L principal components of

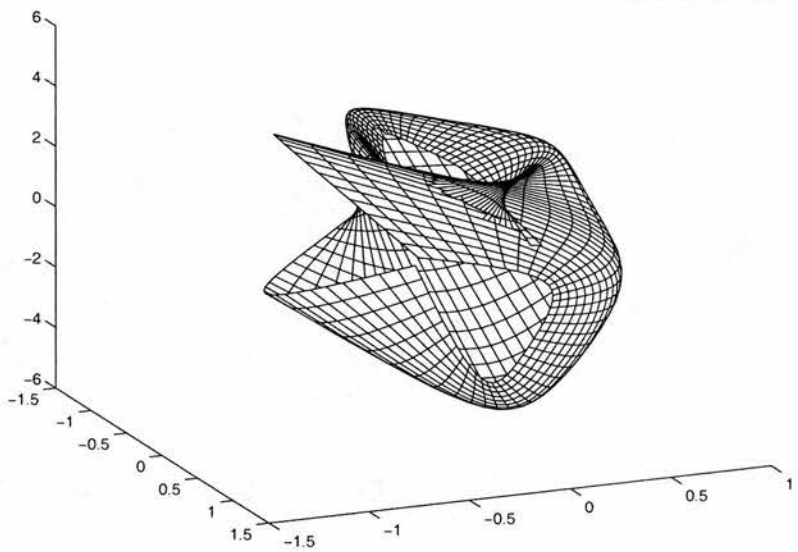

Figure 2.6: A sample manifold from more flexible GTM. 1600 latent points in two dimensions are passed through 16 Gaussian basis functions with means 0.5 standard deviations apart, and mapped via a random matrix into data space. Neighbouring means in latent space are connected by vertices.

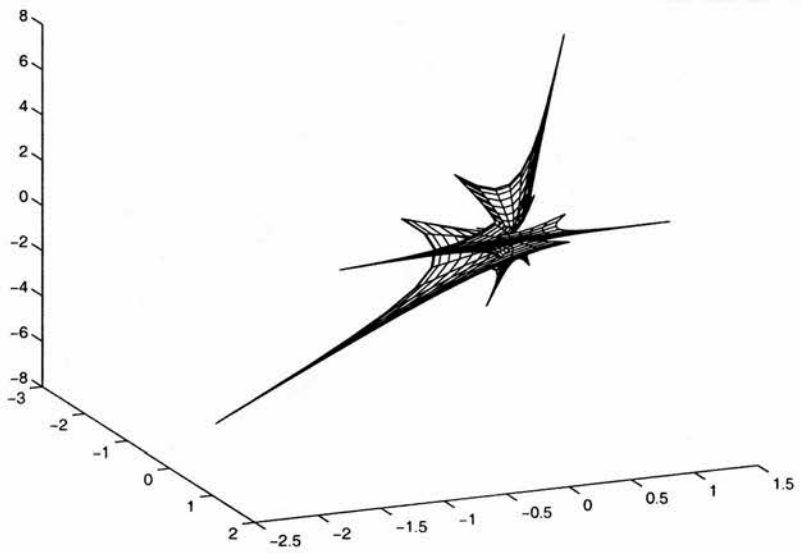

Figure 2.7: A sample manifold from an under-constrained GTM. 1600 latent points in two dimensions are passed through 16 Gaussian basis functions with means 0.25 standard deviations apart, and mapped randomly into data space. Neighbouring means in latent space are connected by vertices.

the data set. This mapping maintains maximum topography since any two points in latent space are certain to reflect nearby points in data space, but may do so by failing to represent curved or non-linear structure in the data. In contrast, a large number of less overlapping basis functions allows a more locally non-linear manifold structure. This mapping allows the manifold to reflect a lot of detail in the data at the expense of topography, since nearby points in latent space are less constrained to represent nearby points in data space. As the basis functions separate, the mapping approaches the mixture of Gaussians model where means are uncoupled. In the figures, only the distance between basis function centres are changed.

2.5.3 Noise and Neural Interpretation

As in PPCA there is a global noise level given by β^{-1}. This reflects the average distance between data points and the map surface. As $\mathbf{W}$ is altered during training the map surface twists and expands to cover the data points more effectively; the closer the map fits the data the smaller β^{-1} becomes. As the variance shrinks, the distribution of responsibilities settles on a small number of latent points forming a localised bump when projected into the latent space.

β may also be understood as a deterministic annealing parameter. Annealing is a method of global optimisation derived from statistical thermodynamics. Standard optimisation techniques, such as EM and neural network competitive learning and gradient descent schemes, are hill-climbing algorithms that at each step move to increasingly probable parameter values[6]. This leads to problems with convergence to local maxima of the likelihood function. In map models, local minima corresponds to twisted, tangled or otherwise suboptimal configurations of the map in data space (Hertz et al., 1991).

Deterministic annealing (DA; Rose et al., 1990; Ueda and Nakano, 1998) is a relaxation method for finding good parameter values by performing sequential optimisations. It also has an interesting neural interpretation with respect to the GTM.

Deterministic annealing starts by convolving (smoothing) the surface of the negative log probability for the parameters to make it convex. This is called a high temperature state, in analogy with annealing in metallurgy. It is then easy to find the minimum. The surface is then gradually deconvolved, or cooled, and EM optimisation is performed

[6]It is well known that supervised neural network objective functions are typically negative log posterior probabilities for the parameters (see Bishop, 1995, ch.11). It is less widely appreciated that the same is true of unsupervised nets (Ritter et al., 1991; Luttrell, 1994).

at each deconvolution step.

In the GTM high temperature is represented by large values of β^{-1}. When the model is given a very broad noise model then the optimal values for $\mathbf{W}$ are straightforward to find – only the large scale features of the data, i.e. its global variance structure, are worth distinguishing from the noise. Consequently the best values for $\mathbf{W}$ map the latent points to a relatively flat manifold reflecting the principle directions of global variance; the large noise variance ensures that any more detail in the data is attributable to noise. The temperature is then decreased slightly by decreasing β^{-1}. Now more local structure must be accounted for by the map because there is less noise, so $\mathbf{W}$ is altered to take into account the newly visible local variance structure. Every time the noise is decreased more of the data's structure is attributed to the mapping, and less to the noise. DA is computationally more expensive then simple optimisation, but ideally, the final values for $\mathbf{W}$ are more nearly globally optimal.

In the neural interpretation of DA and the GTM we identify the latent points with neurons and β with the width of their tuning profile in the 'data space' determined by the structure of their dendritic trees. The flexibility and thus the representational capacity of the neural map is determined by lateral connectivity in the neural sheet. This is analogous to the GTM's fixed basis functions, though the basis functions themselves have no direct analogue. Large values of β give broad and imprecise tuning curves and small values give highly localist coding, where only one neuron represents a particular area of input space. Following the generative turn we can identify feedforward processing as the computation of a posterior probability distribution across the map a range of activity levels reflected in firing rates across the neurons (Oram et al., 1998)

Three situations where we might assume large variance are during development, under processing uncertainty, and during task (re)learning. During development the representational capacities of the tissue are still being determined so it would make sense for the tissue not to commit to any particular detailed structure which would require a very flexible map (Graepel et al., 1997). This would correspond to a simple, i.e. nearly linear, map with much sensory information attributed to ambient noise. When processing ambiguous stimuli the map is already formed on the basis of previous data, so the mapping itself may be detailed, i.e fairly non-linear. However the level of uncertainty about the correct representation is still affected by β^{-1}. Large values for the variance lead to more uncertainty in the posterior distribution. This approach is

explored elsewhere with reference to aphasia (Lowe and Blumstein, 2000). When re-learning a task, perhaps prompted by a sudden drop in the data probability $p(\mathbf{T} \mid \mathbf{W}, \beta)$ due to a shift in the structure of the environment, an increase in the variance of the noise model would allow the map to reconfigure according to the new data. Krekelberg and Taylor (1997) have suggested neural analogues to the variance that behave in the appropriate way to restart the annealing process.

Elements of this neural interpretation of the GTM, and more generally of explicitly latent variable approaches to understanding neural coding are scattered across the literature. Hopefully, stating them for the GTM as a unified application of the generative turn will stimulate more (see Oram et al., 1998; Zemel et al., 1998; Rao and Ballard, 1997, for further examples).

The next section reviews other topographic map models and shows them to be special cases of, or approximations to, a formulation of the GTM based on Gaussian Processes.

2.6 Other Neural Network Models

2.6.1 Soft Topographic Vector Quantisation

Graepel et al. (1997, 1998) have developed the Soft Topographic Vector Quantiser (STVQ) as a general clustering model derived from statistical thermodynamics. Graepel et al. consider the problem of sending information down a noisy communication channel. The standard solution to this ubiquitous information-theoretic problem is vector quantisation (Gray, 1984).

In vector quantisation, a sender and receiver each agree on M codewords or means, indexed by $1 \ldots M$, with which to code the data. For each datum $\mathbf{t}$ to be sent, the sender finds the nearest of the means and sends its index, say i, over the channel. The receiver then reconstructs $\mathbf{t}$ on the basis of the index. When there are fewer means than data points, some distortion is inevitable. The aim of vector quantisation algorithms is to choose positions for the means in data space so that the receiver's reconstruction error is minimised. When, in addition to the distortion due to quantisation, there is structured channel noise that scrambles the indices as they are transmitted, the vector quantisation problem is harder because the means should be positioned to minimise the effects of distortion and of channel noise simultaneously.

The channel noise is represented as an $M \times M$ matrix $\mathbf{H}$, where $\mathbf{H}_{ij}$ represents the probability that index i will be mistakenly transmitted as j in the channel. To get a topographic map the channel noise must be structured so that indices will flip to indices that are nearby on the number line. for example, 3 should be more likely to be mistakenly received as 4 than as 1. To represent a L-dimensional latent space indices must be L-dimensional vectors. for example, if $L = 2$ then $[1, 2]$ should be more likely to be mistakenly received as $[2, 2]$ than as $[10, 6]$.

Graeppel *et al.* represent the index corresponding to each mean $\boldsymbol{\mu}_i$ as $\mathbf{x}_i$, and make scrambling probability depend on distance in the set of indices,

$$\mathbf{H}_{ij} \propto \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\lambda^2}\right) \tag{2.24}$$

where $\sum_j \mathbf{H}_{ij} = 1$ and λ sets the scale of the channel noise.

The STVQ derivation starts by defining a cost function taking into account both distortion and structured channel noise described by $\mathbf{H}$ that is assumed to be known in advance.

$$\mathrm{E}(\{\omega\}, \{\boldsymbol{\mu}\}) = \sum_k^N \sum_i^M \omega_{ki} \, \mathrm{E}_i(\mathbf{t}_k, \{\boldsymbol{\mu}\}) \tag{2.25}$$

$$\mathrm{E}_i(\mathbf{t}_k, \{\boldsymbol{\mu}\}) = \frac{1}{2} \sum_j^M \mathbf{H}_{ij} \| \mathbf{t}_k - \boldsymbol{\mu}_j \|^2 \tag{2.26}$$

As before the latent indicator variable $\omega_{ki} = 1$ if $\mathbf{t}_k$ is assigned to (or generated by) $\boldsymbol{\mu}_i$, and zero otherwise. Intuitively, each datum is assigned to a mean by $\boldsymbol{\omega}$ and the index, say $\mathbf{x}_i$, is passed across a noisy channel, so the cost of the assignment $\mathrm{E}_i(\mathbf{t}_k, \{\boldsymbol{\mu}\})$ depends not only on the error that results from reconstructing $\mathbf{t}_k$ as $\boldsymbol{\mu}_i$ but also on the error that would result if $\mathbf{t}_k$ were reconstructed as $\boldsymbol{\mu}_j$ weighted by the probability $\mathbf{H}_{ij}$ of this scramble occurring. To minimise this cost function it is necessary that means with similar indices deal with nearby regions of the data space, otherwise the reconstruction cost when the index is randomly flipped will be too high. On the other hand, means with dissimilar indices need not deal with nearby regions of data space because the probability of the index being flipped accidentally is low.

As the notation suggests $\mathbf{x}_1 \ldots \mathbf{x}_M$ play the same role in STVQ as the latent sample points in the GTM. Although there is no explicit representation of a latent space in STVQ, λ effectively controls the flexibility of the model, in the same way as the basis function number and spacing in the GTM because large values of λ make coping with

the structured channel noise more important than finding a good quantisation of the data. Conversely, as $\lambda \rightarrow 0$ the pure vector quantisation model is recovered because channel noise no longer affects the transmission process.

The probability distribution corresponding to Equation 2.25 is considered to be unknown, so Graepel et al. infer the Gibbs distribution as the maximum entropy distribution for a given average of the cost function,

$$p(\{\boldsymbol{\omega}\}, \{\boldsymbol{\mu}\}) = Z^{-1}_{(\{\boldsymbol{\omega}\},\{\boldsymbol{\mu}\})} \exp(-\beta \mathrm{E}(\{\boldsymbol{\omega}\}, \{\boldsymbol{\mu}\})) \tag{2.27}$$

where the inverse temperature parameter β is a Lagrange multiplier determined by the average of the cost function. The final distribution over means is obtained by marginalisation

$$p(\{\boldsymbol{\mu}\}) = \prod_k^N \sum_i^M Z^{-1}_{(\{\boldsymbol{\mu}\})} \exp(-\beta \mathrm{E}_i(\mathbf{t}_k, \{\boldsymbol{\mu}\})) \tag{2.28}$$

(see Graepel et al., 1997 for details). This formulation should be compared with Equation 2.6. Equation 2.28 represents the full probabilistic model underlying STVQ.

For a given value of β, fixed point solutions for the means are given by

$$\mu_i = \frac{\sum_k^N \mathbf{t}_k \sum_j^M \mathbf{H}_{ij}\, p_{\mathrm{STVQ}}(\omega_{ki} \mid \mathbf{t}_k, \{\boldsymbol{\mu}\})}{\sum_k^N \sum_j^M \mathbf{H}_{ij} p_{\mathrm{STVQ}}(\omega_{ki} \mid \mathbf{t}_k, \{\boldsymbol{\mu}\})} \tag{2.29}$$

where

$$p_{\mathrm{STVQ}}(\omega_{ki} \mid \mathbf{t}_k, \{\boldsymbol{\mu}\}) = \frac{\exp(-\beta \mathrm{E}_i(\mathbf{t}_k, \{\boldsymbol{\mu}\}))}{\sum_j^M \exp(-\beta \mathrm{E}_j(\mathbf{t}_k, \{\boldsymbol{\mu}\}))} \tag{2.30}$$

$$= p(\mathbf{x}_i \mid \mathbf{t}_k \{\boldsymbol{\mu}\}) \tag{2.31}$$

This expression is a posterior distribution over points in the latent space for the point $\mathbf{t}_k$. Iterating between Equations 2.31 and 2.29 constitutes an Expectation Maximisation algorithm that raises the probability of the mean positions at each step until a local maximum is reached.

β is an annealing parameter. Initially, small values of β spread the posterior distribution over a large number of indices so most means, and therefore many data points, contribute to the updated mean positions. As β is raised the posterior distribution focuses on a smaller number of means, so only the few data near those means contribute to new mean positions.

As in the GTM, a map of the data is obtained by computing $\text{argmax}(a)\, p(\mathbf{x}_a \mid \mathbf{t}_k\{\boldsymbol{\mu}\})$ for each data point k and plotting them as if they were a latent space. It is less clear that taking a posterior mean makes probabilistic sense because the STVQ has no explicit latent space, only index-perturbing channel noise. Unlike the GTM, there is no explicit integral for which the mixture model formulation of Equation 2.28 provides a Monte-Carlo sample (though see Luttrell, 1994).

Self-Organizing Map

The batch Self-Organizing Map (Kohonen, 1982, 1995) is special case of the STVQ when $\beta = \infty$ and Equation 2.31 is replaced by a nearest neighbour rule that ignores channel noise,

$$p_{\text{SOM}}(\omega_{ki} \mid \mathbf{t}_k\{\boldsymbol{\mu}\}) = \delta_{ka} \tag{2.32}$$

$$a = \text{argmin}(a) \parallel \mathbf{x}_k - \boldsymbol{\mu}_a \parallel^2 \tag{2.33}$$

Unfortunately this is not quite the standard SOM; since there is no noise model to alter, the training process must alter λ. This means that the SOM does not have unified error function, probabilistic or otherwise (Erwin et al., 1992). In contrast, by decoupling the structure of the channel noise, controlled by λ, from the trajectory of the parameter optimisation process controlled by β, the flexibility of the map can be held constant as the parameters are annealed. Two principle advantages of this separation are that the channel noise can take any form without hampering self-organization, and that it is possible to construct a neural interpretation of the model similar to that of the GTM.

The neural interpretation of the SOM traditionally identifies the neighbourhood function described by $\mathbf{H}$ as representing excitatory lateral connections in the cortex. However, the range of lateral connectivity only changes substantially during early development (Kandel et al., 1991), which presents an important disanalogy with the SOM.

Krekelberg and Taylor (1997) have suggested that nitric oxide release from neurons might form an appropriate replacement for lateral connectivity. Since nitric oxide is a gas, it has approximately Gaussian diffusion properties, as described in Equation 2.24. However, although Krekelberg and Taylor present a reasonable computational explanation for how nitric oxide release could decrease appropriately as required in the SOM, they admit that empirical studies are equivocal. In contrast, the $\mathbf{H}$ in the STVQ can still be construed as representing fixed lateral connectivity. The annealing parameter

β then reflects either ambient neuronal noise or a measure of uncertainty dependent on the processing task. Indeed, there is no reason to restrict the neighbourhood function to focal excitation; more neurologically motivated patterns of lateral connectivity are possible, without disrupting self-organization.

2.6.2 The Elastic Net

The Elastic Net was first used to solve the Travelling Salesman problem in combinatorial optimisation. For the standard problem $L = 1$ and the data space is a two-dimensional map of cities. The optimisation problem is then equivalent to finding the shortest line that passes through N points $\mathbf{T}$ in two-dimensions. The Elastic Net defines a line by specifying control points given by M Normal distributions with means $\{\boldsymbol{\mu}\} = \boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_M$ in the data space. The prior is uniform in one dimension, so the prior probabilities of each latent point is $1/M$.

If the generating distribution is known to be i then the probability of generating a city $\mathbf{t}$ is Normal with variance κ^2,

$$p(\mathbf{t} \mid \boldsymbol{\mu}_i, \kappa^2) \quad = \quad \mathcal{N}(\boldsymbol{\mu}, \kappa^2 \mathbf{I}). \tag{2.34}$$

κ is an annealing parameter that is slowly reduced during the training process. This expression should be compared with Equations 2.12 and 2.13 (the indicator variable $\boldsymbol{\omega}$ has been suppressed).

The identities of the distributions that generated each of N data points $\mathbf{T}$ are unobserved so the probability of the data set is a product of mixture models,

$$p(\mathbf{T} \mid \{\boldsymbol{\mu}\}, \kappa^2) \quad = \quad \prod_{k=1}^{N} \frac{1}{M} \sum_{i=1}^{M} p(\mathbf{t} \mid \boldsymbol{\mu}_i, \kappa^2) \tag{2.35}$$

This expression should be compared with Equations 2.6 and 2.28.

In the GTM the position of each mean $\boldsymbol{\mu}_i = \mathbf{y}(\mathbf{x}_i; \mathbf{W})$ in the data space is controlled by the relation between $\mathbf{x}_i$ and its neighbours in the latent space; nearby points tend to generate similar values of $\mathbf{y}$. Consequently there is no need to constrain the positions of means in the data space directly. In contrast, the elastic net constrains the positions of means directly in the data space based on their indices. In the Elastic Net the mean

positions are controlled by an improper (spline) prior[7] (Durbin et al., 1989)

$$p(\{\boldsymbol{\mu}\} \mid \kappa^2) \propto \prod_{i=1}^{M} \exp\left(-\frac{\beta}{\alpha\kappa}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1}\|^2\right) \tag{2.36}$$

This prior encourages each mean to be near its neighbours. Following the original papers α and β are constants controlling the relative importance of the data and the prior respectively (Durbin et al., 1989), though in probabilistic formulation only their ratio is relevant. The ratio controls the flexibility of the resulting map by balancing the influence of the prior (Equation 2.36) and the likelihood (Equation 2.35).

For fixed values of κ the posterior distribution of means is given by Bayes theorem:

$$p(\{\boldsymbol{\mu}\} \mid \mathbf{T}, \kappa^2) \propto p(\mathbf{T} \mid \{\boldsymbol{\mu}\}, \kappa^2)\, p(\{\boldsymbol{\mu}\} \mid \kappa^2) \tag{2.37}$$

The right hand side of Equation 2.37 is treated as a function of $\{\boldsymbol{\mu}\}$ and maximised to obtain the Maximum A Posteriori values for the means.

The annealing parameter κ controls the trade off between data fit and topography during training. At high values of κ the positions of the means are dominated by the prior since $\beta/\alpha\kappa \gg \kappa^2$. Consequently the means have perfect topographic ordering at the expense of any fit to the data. At low values of κ the noise model dominates, since $\kappa^2 \gg \beta/\alpha\kappa$, so data fit is much more important than topography. Like the batch SOM, the training process starts at high κ, finds a maximum of $p(\{\boldsymbol{\mu}\} \mid \mathbf{T}, \kappa^2)$, lowers κ slightly and repeats the process.

Like the GTM, the Elastic Net can be understood as a constrained mixture model approximation to the intractable integrals involved in fitting a one-dimensional manifold to two-dimensional data. The continuous nature of the manifold that is assumed to characterize the data generation mechanism is more explicit in Hastie and Stützle's (1989) related Principal Curve model. The annealing parameter is necessary to attain reasonable minima in the parameter optimisation process.

In all the models discussed here, local minima correspond to twists and tangles in the resulting maps. Often twists and abrupt transitions in the positions of means with similar indices are inevitable and biologically interesting. Goodhill (1999) uses the Elastic Net for modelling the formation of ocular dominance columns (Goodhill and Willshaw, 1994) and orientation columns (Goodhill and Cimponeriu, 2000) in V1.

[7] All means can be multiplied by an arbitrary constant and still have the same distribution because the prior depends only on their differences, not their absolute values (C. K. I. Williams, personal comm.)

In the ocular dominance context the optimal (and observed) map structure alternates regularly between eye preferences. This occurs because a one dimensional manifold is attempting to fit essentially two-dimensional data. Ocular dominance columns are also an interesting case where the structure of neural processing reflected in dominance columns is strongly affected by the inability of the underlying model to match the structure of the data. Orientation columns also show abrupt discontinuities called pinwheels that are inevitable because the angular structure of the data necessarily conflicts with the desire for smooth topographic representation (Goodhill et al., 1997; Goodhill and Richards, 1999).

2.6.3 Regularisation and Constraint

The Elastic Net is a regularised mixture model because the neighbouring means are encouraged rather than constrained to stay close to one another by the prior. The GTM on the other hand is a constrained mixture model. The difference is that a constrained mixture that is inflexible, e.g. a GTM with few basis functions, will be unable to represent certain manifolds accurately whereas a heavily regularised mixture, e.g. an Elastic Net with large β/α, is able to represent any data structure at the cost of a very low posterior probability.

The GTM can be converted into a regularised model by replacing the generalised linear mapping in Equation 2.18 by a Gaussian Process (Williams and Rasmussen, 1996; Rasmussen, 1996). Theoretically this corresponds to taking the number of basis functions to infinity while keeping the weights appropriately scaled (Williams, 1998). Since the weight matrix $\mathbf{W}$ has been integrated out, only regularisation parameters remain.

Gaussian Process GTM

Gaussian Process formulation of the GTM places a prior distribution over mapping from latent to data space. For D-dimensional data the Gaussian Process defines a

prior over functions from latent space points onto each dimension of the data space.

$$p(\{\boldsymbol{\mu}\} \mid \nu, \lambda) = \prod_{d=1}^{D} p(\boldsymbol{\mu}_{(d)} \mid \nu, \lambda) \tag{2.38}$$

$$= \prod_{d=1}^{D} Z^{-1}_{(\nu,\lambda)} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}_{(d)}\mathbf{C}^{-1}\boldsymbol{\mu}_{(d)}\right) \tag{2.39}$$

$$\mathbf{C}_{ij} = \nu \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\lambda^2}\right) \tag{2.40}$$

where $\boldsymbol{\mu}_{(d)}$ is a column vector containing the dth dimension in each of the M means and Z^{-1} is a normalising constant for a product of D one-dimensional Gaussians. Maps sampled from this prior have (on average) zero mean and their flexibility depends on λ. Applying Bayes theorem generates a posterior distribution over functions into the data space, and we take the expectation of this distribution to define the topographic mapping.

$\mathbf{C}$ is a matrix derived from the covariance function $C(\mathbf{x}, \mathbf{x}')$ which maps any two vectors onto a Real number representing how similar $\mathbf{y}$ and $\mathbf{y}'$ should be. The formulae above create a topographic map because they ensure that nearby latent space points map to similar location in the data space. Since λ controls the smoothness of the function class. Small values of λ mean that nearby points in latent space can be mapped to quite different positions in the data space; this leads to a very flexible mapping at the expense of topography. Large values of λ lead to very smooth mappings that maintain topography at the expense of fitting the data. ν represents the expected vertical scale of the mapping into data space, similarly to α in the original GTM.

The GP formulation of the GTM is more interpretable than the original because the flexibility of the map is controlled entirely by λ rather than by a set of basis functions and their spacing. More generally Gaussian Processes are an optimal Bayesian solution to the problem of non-linear regression (Williams, 1998) and a potential replacement for multilayer neural networks MacKay (1998). The cost of this optimality for the GTM is that the associated EM algorithm requires inverting the $M \times M$ matrix every cycle, in contrast to the original formulation which inverts a matrix of dimension given by the number of basis functions. From this perspective the GP formulation serves best as a motivation for the original formulation. For the size of problems tackled below, the original formulation is used. To see how they are related we can formulate the original GTM as a Gaussian Process.

The covariance matrix associated with the original formulation of the GTM with B evenly spaced Gaussian basis functions with width l is (MacKay, 1998)

$$\mathbf{C}_{ij} = \alpha^{-1} \sum_{b=1}^{B} \exp\left(-\frac{\| \mathbf{x}_i - \boldsymbol{\mu}_b \|}{2l^2}\right) \exp\left(-\frac{\| \mathbf{x}_j - \boldsymbol{\mu}_b \|}{2l^2}\right). \tag{2.41}$$

Since this matrix is a sum of only B outer products it will typically be rank deficient, but it does serve to show how the basis function parameters relate to the smoothness of the map. As $B \to \infty$ this expression converges to Equation 2.40 with $\nu = \alpha^{-1}$ and $\lambda = l$. Thus the original GTM is a finite basis function approximation to the GP formulation.

A different approximation to the GP GTM is Atsugi's (1998) Bayesian Self-Organizing Map. The prior over means is replaced by a spline regulariser. For a one dimensional manifold $\mu_i = y(\mathbf{x}_i)$ the spline prior

$$\log p(\{\mu\}) \propto -\frac{1}{2} \int \left[D^2 y(\mathbf{x})\right]^2 d\mathbf{x} \tag{2.42}$$

penalises overly flexible mapping by using the differential operator $D^p : y(\mathbf{x}) \mapsto y''(\mathbf{x})$ to constrain second derivatives. With a zero mean and extra terms to tie down first derivatives this prior corresponds to the GP GTM covariance function with $\mathbf{C} = [D^2]^\mathsf{T} D^2$ (MacKay, 1998). The resulting model is less probabilistically interpretable but more computationally tractable than the GP GTM. It should be noted that very similar discretized second derivative terms appear in the Elastic Net which suggests that the Bayesian Self-organizing Map is essentially an Elastic Net augmented with more statistical machinery. The fact that splines are an approximation to the covariance function above also shows the Elastic Net to also be an approximation to the GP GTM.

2.7 Making Maps

From a mathematical perspective we have seen that there are only a few ways to make topographic maps and that they tend to be approximations or special cases of each other. All the topographic map models discussed above are types of Gaussian mixture, and most mixtures can be construed as Monte-Carlo approximations to the intractable integrals generated by a full probabilistic model. In this sense the full GP formulation of the GTM can be taken as the primary, though computationally demanding, statistical map model. The GP GTM assumes a continuous latent space $\mathbf{x}$, a flexible parametric

mapping into a continuous data space $\mathbf{y}(\mathbf{x})$ and a Gaussian noise model with variance β^{-1}. The smoothness of the mapping controls the balance between topography and fit to the data.

The original SOM has a similar structure although it contains no noise model, creates the mapping with kernel regression (Mulier and Cherkassky, 1995) and systematically decreases the flexibility of the mapping over training. STVQ reformulates the original SOM as the solution to the information theoretic problem of optimal encoding over a noisy channel. Since Information theory and Bayesian statistics are intimately linked we might expect to find strong similarities between the information theory solution of STVQ and the generative model solution of the GP GTM. We have seen above that the $\mathbf{H}$, which describes the structure of the channel noise plays the same role as the covariance function $\mathbf{C}$, and that we again recover a constrained mixture model as the optimal model, with SOM as a computationally cheaper approximation. The Elastic Net, and its recent development the Bayesian Self-organizing Map are also a mixture model with a spline prior on the mapping in data space. The previous section showed how splines are an approximation to a full Gaussian Process prior over functions and described how spline priors relate to their associated covariance functions. The original GTM was also shown to be an approximation to the GP GTM with a finite basis approximation to the full covariance function.

2.8 Conclusion

The GTM is a generative latent variable model that extracts a low-dimensional manifold from high-dimensional data. It is also a neural map model. Thinking of neural maps in a latent variable framework suggests that activation levels on a map surface due to a particular input are to be interpreted as posterior probabilities of the position of that input in the low-dimensional structure that describes the variance structure of all possible data. The map embodies the statistical assumption that although the input may appear to be of high-dimensionality, its intrinsic dimensionality is in fact low. When correct, this assumption helps the map creates highly informative two-dimensional projections of the data.

The GTM is a non-linear extension of Factor Analysis. The cost of this non-linearity is that, unlike the linear models considered at the beginning of the chapter, the model specifies an analytically intractable marginal distribution over the data. When this is

approximated by Monte-Carlo methods, the GTM is a constrained mixture model with the same architecture as a radial basis function network. The GTM can be generalised by using a Gaussian Process to map latent points into the data space. This places the GTM in a Bayesian regression framework and allows the smoothness and flexibility of the resulting map to be specified straightforwardly.

The GP formulation of the GTM constitutes a complete probabilistic model of topographic mapping. We have shown that several widely-used alternative models — STVQ, Kohonen's self-organizing map, the Elastic Net, the Bayesian Self-organizing Map, and the original GTM — are special cases of or approximations to the GP formulation.

Lastly we have presented some brief biological interpretations of the key elements of the model. The most important of these is that activity levels in neural tissue should be interpreted as distributions of posterior probability over a latent variable that reflects a particular processing problem. This view is a simple consequence of the generative turn in neural network research and is gaining popularity among population coding researchers (e.g Zemel et al., 1998) and in the vision literature (e.g. Knill and Richards, 1996). The following chapters are an attempt to extend this style of modelling to cover lexical semantic representations.

Chapter 3

Semantic Memory and Priming

Reaction times for naming and performing lexical decision on a target word are faster if subjects have been presented briefly with a related prime word up to several hundred milliseconds before. Semantic priming refers to speeded reaction times due to a semantic connection between the prime and target. Semantic priming is an important method for inferring the structure of semantic memory because it generates measures of semantic relatedness that are not under subjects' conscious control. Consequently, predicting priming effects is an important test of any theory of lexical semantic representation.

Priming phenomena

Priming has been reported for a wide range of lexical relations (Neely, 1991, for a review). We review a few of the most theoretically important varieties, before considering their relevance to memory models.

- If one word is frequently generated in a free association task after presentation of another, e.g. "bed" and "pan", then they are said to be associatively related (see e.g. Deese, 1965); associatively related words generate priming effects.

- Priming occurs between words that are intuitively related in meaning. Typically words are members of the same taxonomic category, e.g. "plate" and "bowl". This is the basic semantic priming effect. Shelton and Martin (1992) have argued that true semantic priming does not occur without association also being present. They attempted to distinguish the effects of association from those of semantic relatedness in a experiment that compared semantically related pairs with those

that were both semantically and associatively related. Facilitation occurred only for the mixed condition.

- The standard account of why "stripes" is named more quickly after a prior presentation of "lion" is that activation of the concept of lion, activates the related concept of tiger, which activates the related concept of stripes. Tiger is said to mediate between the original concepts, generating mediated priming effects. Mediated priming effects are typically smaller than direct semantic effects, and are unreliable across experimental paradigms (de Groot, 1983; Balota and Lorch, 1986).

- Ratcliff and McKoon (McKoon and Ratcliff, 1992) showed that the amount of facilitation generated by prior presentation of related words can be controlled quantitatively by choosing prime words according to a statistical method (Church and Hanks, 1990). The method was originally designed to measure strength of collocation between word pairs for lexicographic applications; larger priming effect size correlates with larger values of the measure for word pairs, generating a graded priming effect.

- Recently Moss and colleagues (Moss et al., 1995) have shown that a wide range of meaning relations generate priming effects, in addition to the standard taxonomic relations. In particular, words from scripts, e.g. "waiter" and "wine", facilitate each other, as do words in instrumental relations, e.g. "rake" and "leaves".

These effects suggest constraints on the architecture and parameterisation of memory models. Accommodating the basic semantic priming effect means that semantically related words (or at least their concepts) should be represented in such a way that they are more easily reachable than words that have no semantic relation. Moss and colleagues' results place more specific constraints on semantic organization. Associative priming effects present similar constraints for associatively related words. Memory models differ in their account of the distinction between associatedness and semantic relatedness (and whether they treat semantic relatedness as distinguishable from associative effects). These differences are discussed below and investigated empirically in the next chapter. Graded and mediated priming express more detailed quantitative requirements for the predictions of memory models. Mediated priming has been put forward as an effect that distinguishes between models that use the concept of activa-

tion in memory and those that do not. We shall see below and in the next chapter that mediated priming cannot play this role.

The psycholinguistic literature makes use of a wide variety of models to explain how priming effects are a consequence of particular architectural assumptions about semantic memory. Below we review two non-statistical memory models, spreading activation and compound cue theory and consider how they deal with the priming data described above. We then consider instances of several styles of neural network and then semantic space models. The ability of each to deal with priming data is reviewed in preparation for the next chapter.

3.1 Non-statistical models

3.1.1 Spreading activation

Collins and Loftus (1975) describe an early instance of a spreading activation model of semantic memory (see also Collins and Quillian, 1969; Anderson, 1983). The model assumes that semantic memory is organized in a taxonomic (and therefore acyclic) inheritance graph with nodes corresponding to concepts; each node in the hierarchy is connected to its parents and children by links representing simple possession and membership relations e.g. `bird` HAS-A `wing` and `cat` IS-A `animal`. The number of connections that must be traversed to get from one node to another gives a measure of the semantic similarity of the associated concepts; distant concepts are less related than nearby ones. For psycholinguistic processing Morton's related Logogen model (Morton, 1979) claims that ease and therefore speed of recognition depends on the level of activation of each word. The resting level is typically proportional to the word's frequency. When a word is processed, activation spreads automatically to nearby concepts in the graph, raising their activation levels above resting levels. However activation decays over distance, which implies a time course for the activation of neighbouring concepts.

Given a suitable graph structure, the spreading activation framework can be extended to cover priming in the following way: Word meanings are identified with concepts, and node activation level determines the length of time it takes to recognise the word corresponding to that meaning. Then, if 'robin' is presented it activates the corresponding node `robin` which spreads activation to `bullfinch` via `bird` (and perhaps by other paths via `beak` etc.). If 'bullfinch' is subsequently presented, then it is

processed more quickly than an unrelated word because of the raised activation of its corresponding node. To get quantitative predictions each link traversal was assumed to take 100 milliseconds (Anderson, 1976, 1983).

3.1.2 Compound Cue Models

Compound Cue models were first presented by Ratcliff and McKoon (Ratcliff and McKoon, 1981, 1988; McKoon and Ratcliff, 1992) as an alternative to the spreading activation framework. In contrast to most spreading activation models they distinguish between the psychological processing model and the semantic memory model on which it depends. The idea is that a compound cue model should be seen in conjunction with an independently motivated memory model. The Search of Associative Memory model (Raaijmakers and Shiffrin, 1981; Gillund and Shiffrin, 1984) is often used.

In a compound cue model there is no activation in semantic memory. Rather, during a testing episode the experimental stimuli are concatenated, along with minimal detail about the context, into a 'cue'. The cue is then compared to each of a set of 'images' of previous episodes in memory. The cue acquires a 'familiarity' value on the basis of how similar it is to previous episodes which determines how easy it is to process. Cue formation and familiarity calculation depend on the specific memory model that implements the compound cue. For example, in SAM semantic memory contains images that are "closely interconnected, relatively unitised permanent sets of features" describing aspects of the experimental situation. Specifically these feature sets represent a) item information, e.g. the spelling of a word, b) contextual information, the context the item occurs in, and c) inter-item links e.g. associative relatedness information. When an image needs to be retrieved a subject assembles a cue set consisting of features of the current situation, e.g. item cues identifying the currently presented stimuli, and context cues. Each cue activates each image in memory with strength controlled by a matrix of cue-image strengths. Retrieval performance depends on familiarity – a weighted sum of the familiarity levels between the cue set and each image in memory. Thus for a single cue the familiarity is a vector sum across the corresponding row or column of the matrix. In recall, each image is sampled according to the proportion of the total familiarity taken by its cue-image strength in the matrix. Thus very familiar cues are easily and quickly generated.

3.1.3 Priming effects

The original Collins and Loftus model can represent different types of semantic relation explicitly because it contains labelled graphical links. However, no implementation of the compound cue model distinguishes relations other than taxonomic. There is also no obvious way to represent the difference between semantic and associative relatedness, although this may not be a problem if they are in fact generated by the same spreading activation mechanism. Graded priming is straightforward for a compound cue model because familiarity is a continuous quantity, like facilitation. Ratcliff and McKoon also suggest that it depends on word co-occurrences, which need not reflect intuitively semantic distinctions (McKoon and Ratcliff, 1998). Spreading activation models can also deal with graded priming in theory, although Moss *et al.* observe that it is not clear from examination of the stimulus materials *why* the words chosen automatically for the graded priming experiment actually facilitate one another. We consider this question in the next chapter.

Mediated priming has been put forward as a crucial test between the two models. If mediated priming really does depend on a mediation process involving activation spreading in memory then only spreading activation models have the appropriate architecture to capture the effect. Compound cue models cannot deal with mediation literally because "tiger" is never presented and so cannot become part of the cue whose familiarity value is computed. Moreover, if activation spreads through an intervening node then it should be much weaker by the time it reaches "stripes", so the prediction from spreading activation theory is that mediated priming effects should be weaker than direct priming effects. This is also consistent with experimental results.

In response to this difficulty for compound cue theory McKoon and Ratcliff have argued that there is in fact no mediation involved in mediated priming. It is simply that "lion" is weakly but directly related to "stripes" in memory. The theory of weak but direct relatedness in mediated priming is tested in the next chapter. According to this theory, mediated priming is a type of graded priming.

3.1.4 Problems with Non-statistical Models

More recent spreading activation models no longer subscribe to the taxonomic claims of Collins and Loftus (Eysenck and Keane, 1995), but the essential idea of spreading activation in a graphical structure where distance between nodes represents semantic

similarity has remained, sometimes augmented by non-automatic processing mechanisms such as post-lexical checking to account for other effects (Neely, 1991). A more general difficulty with spreading activation models is that it is very difficult to construct the necessary graph structure in the correct way, and, independent of the method of construction, the abstract notion of spreading activation also gives rise to false predictions.

Spreading activation models should predict that the less related one word is to another, the *later* its priming effect onset; activation takes longer to get to words that are many links away. Lorch (1992) and Masson (1991) have pointed out that the onset of priming is not later for less related words, it is only that the asymptotic priming effect is smaller. Subsequent models were then changed; the original 100 msec. per link estimate was revised down to 5-10 msec., and the difference in priming effect was attributed to different rates of growth of activation at near and distant nodes. This rather unprincipled change highlights the fact that the central explanatory mechanism in spreading activation models contains a rather important free parameter: the decay rate of spreading activation. That this is a problem is most clearly seen in models of the effect of an intervening string.

In this paradigm a string intervenes between a prime and an associatively related target during the naming task. The string may be a real but unrelated word, or a neutral stimulus (e.g. XXXX). A natural prediction for a spreading activation model might be that the intervening prime should make no difference to any priming effect – activation still spreads from the prime despite the independent processing of the intervening string. Experimentally, if the string is a word the priming effect is reduced whereas if the string is a neutral stimulus then priming is retained.

On closer inspection it is clear that if the decay rate of spreading activation is a free parameter, almost any pattern of priming can be realized. For example, on the assumption that the neutral prime does not activate any node, a short decay rate will remove the priming effect for neutral and word strings and a long decay rate will maintain priming over many intervening words.

Compound cue models suffer the same construction difficulties as spreading activation models. Their matrix structure is simpler, but almost all the parameters that determine the amount of familiarity generated by a word pair are essentially free, or determined by other unobservable variables, e.g the amount of rehearsal each item in

a list gets.

Until there is a principled way to estimate the decay rate of activation from empirical data, or the matrix parameters in the compound cue model, both models remain useful analogies rather than practical quantitative models. Indeterminacy in traditional non-statistical models have motivated more statistical approaches. We review neural network models and semantic spaces next.

3.2 Statistical Models

3.2.1 Neural Network Models

Like the models discussed above, neural networks are also typically over-parameterised and of very general architecture. The crucial difference, however, is that there exist general purpose algorithms for determining parameter values from empirical data.

Networks used in psycholinguistics form three broad model classes: associative networks (Masson, 1991, 1995), feedforward multilayer perceptrons (Bullinaria and Huckle, 1997) and recurrent networks (Plaut et al., 1994; Plaut, 1995; Moss et al., 1994; Christiansen and Chater, 1999). We examine examples from each class and their ability to deal with the types of semantic priming described above in the next section.

3.2.2 Associative networks

Masson (1991, 1995) uses a Hopfield network (Hopfield, 1982) as a model of semantic memory. Each word is represented by a two part binary vector, $[\mathbf{t}, \mathbf{x}]^\mathsf{T} = [t_1 \ldots t_D, x_1 \ldots x_L]^\mathsf{T}$ where $\mathbf{t}$ is a (randomly generated) representation of the formal properties of the word, e.g. its phonology, and $\mathbf{x}$ represents its semantics. There are then $D + L$ units in the network, and each binary vector is a rudimentary lexical entry. Two words are similar in meaning when the semantic elements in their vectors are close in Hamming distance. Although Masson does not suggest how the latter vector elements are to be generated they have a natural interpretation as semantic features.

Masson's model is a network in which each neuron is connected to all the other neurons. A linear mapping describes the dynamics of the input unit values:

$$[\mathbf{t}, \mathbf{x}]^\mathsf{T}_{(t+1)} = \mathrm{sgn}\big(\mathbf{W}[\mathbf{t}, \mathbf{x}]^\mathsf{T}_{(t)}\big) \tag{3.1}$$

where the sign function ensures that the output is always a binary vector. Equation 3.1

is an instantaneous formulation where all units are changed together. Masson's more neurally-inspired scheme updates one unit at a time, by choosing randomly a row of $\mathbf{W}$, say i, and computing its dot product with the input to produce a new value for unit i. The asynchronous updating rule creates positively skewed distributions of settling times similar to those found in reaction times.

When $[\mathbf{t}, \mathbf{x}]^\mathsf{T}$ has elements of the wrong sign, the recursion described in Equation 3.1 corrects the mistakes and recovers the correct values in a time that depends on the Hamming distance between the initial values and the original vector. This behaviour defines the Hopfield model as an associative memory. When the input vector consists only of the non-semantic vector elements, $[\mathbf{t}, 0]^\mathsf{T}$, then the settling time to $[\mathbf{t}, \mathbf{x}]^\mathsf{T}$ can be taken as a lexical decision time. To model priming between words i and j the network is presented first with $[\mathbf{t}_i, 0]^\mathsf{T}$ and begins to settle toward $[\mathbf{t}_i, \mathbf{x}_i]_i^\mathsf{T}$. The units are then presented with $[\mathbf{t}_j, 0]^\mathsf{T}$, and the system moves toward $[\mathbf{t}_j, \mathbf{x}_j]^\mathsf{T}$. Since the non-semantic vector elements are randomly chosen, the time taken to reach $[\mathbf{t}_j, \mathbf{x}_j]^\mathsf{T}$ depends on the Hamming distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Semantically related words have more bits in common in their semantic vectors so resettling is faster for related words, than unrelated words.

Masson's model is formulated as lexical decision mapping: $\mathbf{t} \rightarrow \mathbf{t}, \mathbf{x}$, taking form and generating meaning. It could, however, equally well be used as a model of 'speech': $\mathbf{x} \rightarrow \mathbf{t}, \mathbf{x}$, by clamping the semantic vectors instead. In this respect it is more natural and flexible than the recurrent networks discussed in the next section.

3.2.3 Priming Effects

Masson's model provides a simple account of the basic semantic priming effect. Semantic relatedness depends on overlapping sets of semantic features so reaction times correspond to Hamming distance in $\mathbf{x}$ space. Graded and mediated priming can also be modelled as progressively *less* overlapping semantic representations. Although it may appear that activation is spread by Equation 3.1, it is not of the same type as a traditional spreading activation model, because the ultimate arbiter of priming is feature overlap. Alternatively the model could be seen as spreading activation in an L-dimensional space, except that it is not activation that spreads between concepts, but lexical entries that spread towards one another.

Ultimately, since Masson's model makes no claims about the structure of $\mathbf{x}$, it is

difficult to know how it would deal with different types of priming relations, or how it might distinguish between association and semantic relatedness, save that every effect must depend on feature overlap.

3.2.4 Problems with associative memory models

Masson's model does give a simple account of priming phenomena. It also has the advantage of a large body of theoretical literature on the capacity and properties of associative networks (see e.g. Hertz et al., 1991). Unfortunately, the principal capacity result for Hopfield networks states that the number of stable states (i.e. recallable lexical entries) in a network of N units with a error tolerance of 5% is bounded by $0.14N$ (Rojas, 1996). This suggests first, that a very large number of neurons need to be available to code for a 30,000 word vocabulary, and second that the more words that are learned the longer the semantic representation for *all* words must be.

Elements of $x_1 \ldots x_L$ may be interpreted as semantic features. They might also be interpreted as indicator variables representing positive lexical association between the word represented by the vector and L other 'context' words. This interpretation suggests a potential unification of associative network models and semantic space. Such a unification is not pursued here because all meaningful interpretations of the semantic vector elements have a common difficulty:

The capacity result above makes strong distributional assumptions about the input vectors; vector elements should be independently distributed. Unfortunately Masson's model of semantic priming works precisely because there *are* non-random correlations between the semantic vector elements of related words, and the capacity of Hopfield networks becomes drastically compromised in proportion to the amount of correlation present (Hertz et al., 1991). But *any* interpretation of the $x_1 \ldots x_L$ that functions as an informative semantic representation must correlate the elements systematically. Stronger semantic relations lead to increasingly smaller active capacities. Perhaps because of this realization, associative network explanations of detailed semantic priming phenomena have not been produced.

3.2.5 Recurrent Networks

Recurrent network models (Plaut et al., 1994; Plaut, 1995; Moss et al., 1994, e.g.) map an orthographic or phonological word representation onto its corresponding semantic

representation: $\mathbf{t} \rightarrow \mathbf{x}$. In addition, recurrent connections couple the current and previous inputs. This forces the mapping to take into account the previous phonological input and semantic output when learning the mapping. In Plaut's networks, reaction times are modelled by the settling times of output units, and are therefore controlled by the evolving output prediction error. As in Masson's model a prime is presented and processed for a small number of recurrent cycles before the target is presented. Settling time also produces positively skewed distributions that depend on the number of shared features in the prime and target semantic vectors.

Unlike Masson's model, there is no practical difficulty in adding structure to $\mathbf{x}$. Semantic vectors were drawn from one of 8 classes. Class prototypes were defined by randomly sampling 100 binary semantic features. Plaut produced both 'high dominance' vectors that differed from their prototypes by a small amount of re-sampling noise, and 'low dominance' vectors that had more frequently re-sampled elements. This produced cluster structure in the 100-dimensional semantic space, and constitutes a simple theory of how word meanings are distributed.

In an attempt to make semantic representations more realistic, Bullinaria and Huckle (1997; 1995) trained a similar cascaded feed-forward network to map orthographic patterns onto target vectors containing 30 principal components of points from a semantic space model corresponding to each input word. The network showed robust semantic priming, although the data were noisy (see Bullinaria and Huckle, 1997, Figure 4.). Averaging over 16 identical networks trained from different random starting positions made the priming effects clearer, and corresponds more directly to the human experimental situation.

In all these networks relatedness between words i and j depended on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, so *if* feature overlap, or position in semantic space is an appropriate operationalisation of semantic relatedness *then* a feedforward net can noisily but successfully use position in semantic space as a target semantic representation. An alternative approach to learning rather than specifying semantic representation is taken by Elman.

Elman (1990; 1991; 1993) trained recurrent networks on a corpus generated from a small grammar that incorporated semantic class structure. Unlike the previous models Elman nets did not learn the mapping: $\mathbf{t} \rightarrow \mathbf{x}$, but rather: $\mathbf{t}_{(t)} \rightarrow \mathbf{x} \rightarrow \mathbf{t}_{(t+1)}$, where $\mathbf{x}$ was learned. After appropriate training ('starting small') the sigmoidal network outputs

generated a posterior probability distribution over possible choices for the next word that closely approximated the optimal distribution given knowledge of the generating grammar. For example if 'man', 'woman' and 'boy' are equally likely expansions for a word of class NP in the grammar, then the values of the output units corresponding to these words closely approximated 1/3 each[1].

Since producing posterior probabilities is an optimal solution to the prediction problem we may be certain that the network was using all the statistical information available. Cluster analysis of the hidden unit activities confirmed that linguistically interpretable structure was in fact represented (Elman, 1993).

From a traditional linguistic viewpoint two kinds of information are predictively useful for the network. The first kind is syntactic: words fall into different part-of-speech classes that constrain their distributional characteristics. For example, the network had to learn how to apply a sub-symbolic version of the rule $S \rightarrow NPVP$ because the grammar ensured that the part-of-speech of the current word was predictive of the part-of-speech of the next. The second kind of information is that words differ in their semantic class. For example, the grammar required that 'eat' took only animate subjects which are only a subset of possible nouns. The important point for theories of lexical semantic representation is that although part-of-speech is a syntactic property of words and animacy is a semantic property of a word's referent, the two types of information are indistinguishable for the network; both are distributional constraints. More generally, they are indistinguishable from a statistical perspective in the sense that a correct grammar of English will contain some distributional constraints that are due to 'syntactic' factors and some that are 'semantic'. In contemporary linguistic theory this observation is increasingly widely appreciated (e.g Pollard and Sag, 1994) and is reflected by the popularity of unification grammar formalisms (Schieber, 1986; Sells, 1985) where there is no representational distinction between phrase structure rules, case marking requirements and semantic constraints on subcategorization structure; each is represented as a constraint on feature structures that allows or prohibits particular unifications.

Elman networks are theoretically attractive because they learn rather than assume representations that reflect the semantic structure of the corpora they are trained on.

[1]This is an approximation because syntactically illegal combinations like NP, NP occur when sentences are abutted but sentence mark-up has been lost. These occurrences provide evidence for small amounts of extraneous probability distributed across the rest of the vocabulary set.

However, while the hidden unit activations provide useful summaries of the distributional properties of a word, they are not accessible to any further processing. Elman himself can show that semantic class structure is present in the hidden unit activations by subjecting them to cluster analysis, but this information is not internally available to the network for any other purpose without a process of representational redescription (Karmiloff-Smith, 1995).

3.2.6 Priming Effects

A common assumption in psycholinguistics is that associative relatedness differs fundamentally from semantic relatedness; association depends on high conditional probability for the target word given the prime. Semantic relatedness in contrast depends on genuinely semantic properties such as shared features, or on predicate structure. A natural way to implement this theory in a recurrent network is to alter the presentation probabilities of words that are intended to be semantically related. Targets that are associated with primes have higher probability of occurring soon after them during training, while semantically related words share features in their semantic representations. This scheme is used by Plaut and Moss and colleagues to generate semantic and associative priming (Plaut et al., 1994; Moss et al., 1994).

Although they have not been used extensively in psycholinguistics, Elman's networks are limited to only producing priming that is consistent with the conditional probability theory of associative priming because semantic features are tuned to produce the *next* word, rather than a related one.

3.2.7 Problems with Recurrent Networks

There are also more general difficulties with recurrent networks as psychological models. First, recurrent networks require computationally intensive training methods, and extremely large numbers of hidden units for realistic-sized models. Consequently, recurrent networks have not been trained on real corpora and are not used as general language models. More problematically, recurrent networks are inherently non-linear regression devices; Plaut and Moss's networks compute $p(\mathbf{x} \mid \mathbf{t})$, the conditional probability of a particular semantic representation given the current input. If the input to a net is the phonological representation of a word and the output is a vector representing its semantics, then the network can perform 'lexical decision' by clamping the input

and examining the output. It should therefore be possible to clamp the outputs and examine the inputs to mimic 'speaking', or to clamp the inputs, infer the meaning, clamp the outputs at that meaning and then examine the inputs to model the naming task. These operations are possible in Masson's model, but there is no appropriate statistical procedure for recurrent networks (Elman nets have similar difficulties predicting the *previous* word, rather than the next one). This is because whereas a recurrent network computes $p(\mathbf{x} \mid \mathbf{t})$, the desired distribution is

$$p(\mathbf{t} \mid \mathbf{x}, D) \quad \propto \quad p(\mathbf{t})\, p(\mathbf{x} \mid \mathbf{t}) \tag{3.2}$$

but we do not know $p(\mathbf{t})$. Worse, the mapping may not even be functional, if there are two words for the same meaning. In theory it would be possible to guess $p(\mathbf{t})$, take a large number of samples and then run the network forward with each of them as input hoping that one of them would produce a value of $\mathbf{x}$ that is close to the one we are interested in conditioning on. This is an expensive and psychologically unmotivated process.

The problem of reversing a feedforward network is an important issue for evolutionary language simulations that model agents as recurrent networks (e.g. Batali, 1998, 2000). Multiple agents must typically 'speak' and 'understand' many strings over many generations. Various statistically questionable proposals have been suggested in that literature (Batali, 1998). Perhaps unsurprisingly these methods appear to introduce significant instability to the evolutionary systems they are part of (Tonkes and Wiles, 2000). There is no reason to think they will work better as psycholinguistic models.

One solution to the reversibility problem is to revert to Masson's model, although it is then impossible to implement the conditional probability theory of associative connection. Another solution is to use a generative model; semantic representations are naturally interpreted as hidden or latent variables, and the observable data of language are sequences of orthographic or phonological elements. Generative models for sequences that might replace recurrent networks include Hidden Markov models, Kalman filters or Stochastic Context-Free Grammars (Roweis and Ghahramani, 1999; Stolcke, 1994, for reviews) or their non-linear counterparts. Reversing the model is a standard part of the training procedure for these models, is statistically well-founded and can be performed exactly. We investigate a generative approach in the next chapter.

3.3 Semantic Space Models

Semantic spaces represent each word as a vector of real numbers that are derived from co-occurrence statistics computed over large corpora. The semantic similarity between words i and j is represented by the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, or by the cosine of the angle between them. Semantic space models are analysed in detail in the next chapter.

Two popular semantic space models in the psychological literature are the HAL model of Lund and colleagues (e.g. Burgess and Lund, 1996; Lund et al., 1995) and the LSA model of Landauer, Dumais and colleagues (Landauer and Dumais, 1997; Berry et al., 1995; Caid et al., 1995). Redington and Chater have also presented semantic space accounts of syntactic and semantic acquisition (Redington and Chater, 1997, 1998) and McDonald and Lowe (McDonald and Lowe, 1998; Lowe and McDonald, 2000) have used semantic space to model priming. The results presented in the next chapter are an extension of these.

The principal advantages of semantic space models is that they may be learned from widely available empirical data, and they thus provide a theory of what makes words similar in meaning at the same time as predicting priming effects.

3.3.1 Priming effects

In semantic space models priming is taken to be proportional to the distance or cosine between word vectors. There is no processing model associated with the space, so all priming effects must be represented by distance or angular structure in space. This makes spaces strong and non-intuitive models of semantic memory. In particular they appear to be inconsistent with the conditional probability theory of associative priming. They also provide one implementation of Ratcliff and McKoon's direct theory of mediated priming, since there is no mediation mechanism available.

Burgess and colleagues have used HAL to model simple semantic priming, but have not successfully modelled associative priming (Burgess and Lund, 1998). They conclude that HAL represents only semantic, and not associative relations. HAL has also been applied to data from mediated priming effects (Livesay and Burgess, 1998b,a). From these experiments Burgess and colleagues conclude that the direct theory of mediated priming must be false.

Unlike HAL, LSA constructs a low-dimensional model of word co-occurrence statistics and calculates distances according to the data's reconstruction according to the

model rather than in the original space directly. In their discussion of the success of LSA, Landauer and colleagues note the importance of dimensionality reduction. They also note a slight bias towards associative rather than semantic relations; 'doctor' was considered to be slightly more similar to 'nurse', its normative associate, than to 'physician', the correct category coordinate synonym. However, the only priming experiments LSA has been applied to have sentence primes and we do not consider them here.

3.3.2 Problems with semantic spaces

Although semantic space models are conceptually simple and can be learnt from data, the principal problem is with performance. HAL is not able to model mediated priming, or associative priming, and has not been applied to graded priming materials or a wide range of semantic relations. Failure to model associative and mediated priming have led Burgess and colleagues to some strong theoretical conclusions about the nature of association and of mediated priming. However, there is more than one way to build a semantic space, so it is presently unclear whether HAL's inabilities derive from its construction or from deep facts about the semantic space framework. If they depend on its construction then an improved space model may capture the missing phenomena, and challenge previous theoretical conclusions. In the next chapter we address these issues empirically with two new semantic space models.

3.4 Conclusion

Neural networks are a significant advance from spreading activation and compound cue models because they require a semantic memory theorist to make explicit assumptions about the structure and contents of memory. In the absence of detailed and realistic spreading activation networks or compound cue systems it is not possible to test specific hypotheses about the representational structure. This is extremely important for models of semantic memory because without specific hypotheses there can be no quantitative comparison to the wealth of semantic priming data available. Currently only very gross architectural aspects have been debated e.g. the existence of spreading activation in the light of mediated priming effects.

However, although recurrent neural networks have an apparently more brain-like structure than traditional models they have seldom generated more than qualitative

accounts of priming phenomena. For example, Plaut has modelled associative and semantic priming in a recurrent network that learns to map phonological representations onto meaning on the basis of input data that has been manipulated according to the conditional probability theory of associative relatedness. However, this success shows only that the combined model is a possible explanation of semantic and associative priming. To make this account convincing it would be necessary to specify phonological and semantic representations appropriate to real experimental stimuli and train the network on these. But this has not been done, principally because we have no idea how to generate semantic representations without building in semantic priming effects. Bullinaria and Huckle have used independently motivated semantic representations (from a semantic space), but the results were not compared to human data. Associative networks are an alternative type of network that generate qualitative priming predictions. However, to work effectively they must make assumptions about the distribution of features in semantic representations that are extremely implausible.

Elman networks are an alternative use of recurrent nets that generate their own semantic representations as a necessary subpart of solving the problem of predicting the next word in a sentence. Apart from the severe computational difficulties that would be involved in scaling the network up to realistic vocabulary sizes, we have seen that Elman networks share a fundamental difficulty with all recurrent and feedforward neural networks. Ultimately they cannot be good psycholinguistic models because they cannot be reversed to generate words from semantic representations.

We have seen that semantic space models solve the problem of generating predictions about priming by using word co-occurrence statistics to create a vector space where distance and angle correspond to semantic relatedness. Semantic space models make detailed and testable predictions about priming effects and can be compared to the results of experimental studies directly. However, they have not yet captured a number of important priming effects. In the next chapter we motivate semantic spaces more fully as models of lexical semantic representation, and present new methods for generating them.

Chapter 4

Semantic Space

Similarity between words can be measured across multiple variables. Some common quantitative measures from psycholinguistics include frequency, orthographic or phonological neighbourhood and associative relatedness. Semantic similarity has previously been an exclusively qualitative variable determined by the experimenter's intuition.

This chapter reviews distributional approaches to meaning and shows how this approach motivates and explains the success of corpus-based semantic space models that represent semantic similarity relations between words geometrically as angles or distances. We then present a general theory of semantic space and use it to analyse two contemporary models, Burgess and colleagues' Hyperspace Analogue to Language (Lund et al., 1995) and Landauer and colleagues' Latent Semantic Analysis (Landauer and Dumais, 1997). We also show how detailed consideration of the task of a semantic space model motivates several new methods for model construction.

4.1 Distributional Approaches to Meaning

Semantic space models can be motivated succinctly with Wittgenstein's injunction "don't look for the meaning, look for the use" (Wittgenstein, 1958) and Firth's observation that "you shall know a word by the company it keeps" (Firth, 1968). If we understand "use" narrowly, as the sorts of linguistic contexts a word typically appears in, and "company" as the words occurring near to it, then we can define a purely distributional sense of semantic similarity: two words are similar to the extent they keep the same (lexical) company. This notion of semantic similarity has led to the idea (Finch, 1993) of a semantic version of the classical replacement test for determining

the syntactic properties of words.

4.1.1 Replacement tests and substitutability

Two words are said to be nouns if they can be substituted into the same sentential contexts while preserving grammaticality (Radford, 1988), e.g. if they can be subcategorized for by verbs and modified by adjectives but cannot modify or be substituted for other verbs. Notice that this definition of what it is to be a noun makes essential use of the concepts of modifier, adjective and verb. However, verbs, modifiers and adjectives are defined similarly, according to their possible patterning with respect to nouns. At first sight this style of explanation looks useless because the part of speech definitions all presuppose that other parts of speech are already defined. The reason this style of explanation is in fact very fruitful is that parts of speech are, like binary branching trees or feature structures, simply latent variables in linguistic theory. Their purpose is to help explain the regularities observed in human judgements of grammaticality (the rest of the explanation is given by ascriptions of grammatical structure). The process of manually determining part of speech is tractable because ascribing a part of speech to a word makes a claim about other words that can be substituted for it while preserving grammaticality; roughly, all nouns may be substituted for one another without making the sentence ungrammatical. Thus some assignments of parts of speech to words allow substitutions that make sentences that we know to be ungrammatical grammatical, and vice versa. The fewer mistaken grammaticality predictions an assignment gives, the better it is.

One interesting consequence of a distributional characterisation of parts of speech is that ascription is a *holistic* enterprise; all parts of speech must be assigned at the same time because none are explanatorily prior.

Another important consequence is that the replacement test provides a way of assigning part of speech without having any advance linguistic knowledge about what nouns, verbs and modifiers *are*, outside of their role as the hidden causes of distributional regularities that govern permissible substitutions between words. It is clear that part of speech ascription can take place in the context of an entirely foreign language (and indeed must have done often in the past). Given only word boundaries, some expectations about the number of parts of speech and enough grammaticality judgements, we may expect part of speech ascription to be tractable, though no doubt difficult, in

extremely impoverished cognitive circumstances[1].

Post-Chomskian syntactic theory has made considerable progress on the assumption that grammaticality judgements are also the data that a theory of grammar must explain, and that such judgements are stable across subjects. Further, recent work in Optimality Theory (Barbosa et al., 1998) and graded grammaticality (Bard et al., 1996; Keller, 1997) shows that even judgements about the *degree* of syntactic well-formedness of a sentence are reliable across subjects. Grammaticality is typically assumed to be a binary property of sentences, but judgements about syntactic well-formedness appear to depend inversely on the number of syntactic rule violations. Just as a statistical part of speech tagger need have no extra-distributional understanding of nouns and verbs, e.g. that nouns often denote objects, to perform effective tagging, so in these experiments there is no reason to believe that subjects are explicitly aware of the rules that are being violated. It is sufficient that they are sensitive to the effects of grammatical structure on the grammaticality of their sentences.

The semantic replacement test assumes that any two words are semantically similar to the extent they may be substituted for one another while preserving sentence meaning. With a sufficient number of different sentences it should be possible to probe the distributional characteristics of any set of words. As in the part of speech test described above, it appears that we must be able to represent meanings in order to check whether sentence meaning is preserved. But again, we will see that this is not the case. The semantic replacement test does, however, entail that speakers are to distinguish degrees of semantic similarity between sentences. Unlike the part of speech test, the semantic replacement test models differences in semantic similarity, with degree of substitutability, an essentially continuous quantity. Substitutability *may* have some unobserved categorical structure (this is assumed by several of the models discussed below) but we do not need to assume this at the outset. This contrasts with the task of inferring parts of speech which are considered by linguists as discrete, albeit unobserved, categories that words can belong to.

It should be noted that a substitutability account of meaning does not need to assume that any two words ever *are* perfectly substitutable with one another. Nor does it need to take a position on the related issue of whether 'true' synonymy is possible.

[1]Note that even in formal linguistics the exact number of speech parts is subject to debate. This indeterminacy is part of the reason there are multiple overlapping tag sets available for corpora (Burnage and Dunlop, 1992).

It is only necessary that the set of contexts available for two words can overlap to a greater or lesser extent. Indeed one of the principal attractions of the theory is that it provides a continuous measure of relatedness without requiring synonymy as a baseline for semantic judgements.

As a first step towards operationalisation we can distinguish two senses of meaning preservation. In the first strict sense, sentence meaning fails to be preserved whenever the replacement word is not an exact synonym. The second looser sense is best explained by example:

We put milk out for the **cat** after supper. (1)
We put milk out for the **dog** after supper. (2)
We put milk out for the **banana** after supper. (3)

Sentence (1) is semantically unexceptional. Sentence (2) has a quite different meaning and is slightly unusual. Sentence (3), like (2) has a quite different meaning (neither sentence provides a good synonym for 'cat'), but is definitely unusual; that is, we would not expect to find it in ordinary language. In the strict sense of meaning preservation, (2) and (3) are equally bad because both are mistranslations of (1). In the looser sense however, (2) is semantically more reasonable. more meaning*ful* than (3) and should be ranked more substitutable because we are much more likely to see (2) than (3) in ordinary language. In the semantic replacement test it is this second sense that is relevant. As hinted at earlier it is more a test of meaningfulness then meaning identity[2].

We don't need to know *why* sentences (2) and (3) are ranked as unusual in this way (presumably because few dogs and no bananas drink milk). The test sentences can only measure how likely we are to see 'dog' than 'banana' in a particular sentence position. That 'dog' is more likely to occur there than 'banana' allows us to conclude that 'dog' is more similar in meaning to 'cat' than to 'banana'. Finally, it is clear that the test allows us to draw this conclusion without knowing anything about cats and dogs. Cruse (1986) presents a semantic theory based on substitutability judgements and other more subtle distributional measures.

4.1.2 Vector space representations of substitutability

The replacement test holds a sentence constant and varies the words, but it can be inverted by considering individual words and varying the sentences that surround them.

[2]Semantic congruence might be a more accurate name, except that it is exceedingly ugly.

For example we might consider the following replacements:

We put milk out for the **cat** after supper. (4)
I have to buy milk for the **cat** and clean up after it. (5)
She takes sandwiches and a **cat** to eat at work. (6)

Sentence (5) is a reasonably meaningful replacement for (4), whereas (6) is not. Sentence (5) is also a possible surrounding context for 'dog', but not for 'banana'. Sentence (6) is, however, a plausible context for 'banana' and inappropriate for 'cat' or 'dog'. As this example suggests, the set of sentences that surround two semantically similar words will overlap more than the set of sentences that surround semantically dissimilar words, so the inverted replacement test gives the same results as the original formulation.

Unfortunately a direct implementation of the inverted test is infeasible. Inverting the test leads to an effectively infinite set of sentences that need to be considered as possible surroundings for a single word. However, if we happen to *have* a large sample of meaningful sentences that contain words of interest, then the inverted test has a considerable advantage: the subject's judgement can be dispensed with. The key to realizing this advantage is to represent sentences in a suitably flexible way.

Consider a set of D context words $b_1 \ldots b_D$. We can represent a sentence as a vector of co-occurrence counts. If the context words are 'we', 'clean', 'eat' and 'milk' then the three sentences above are

[1 0 0 1] (4')
[0 1 0 1] (5')
[0 0 1 0] (6')

In a large enough sample of text a word will occur in many of its possible meaningful sentential contexts and each context will generate a vector. The relative frequency that particular vectors occur will be proportional to the semantic reasonableness of the corresponding sentences.

As an example, assume that we have seen sentence (4) surrounding the word 'cat', sentence (5) with 'dog' and sentence (6) with 'banana' in a large corpus of naturally occurring language. Sentence (5) could also have occurred with 'cat' but in this hypothetical corpus it did not. To get a compact representation of the distributional

properties of 'cat', 'dog' and 'banana' we combine the elements of the vectors for each sentence context of each word, e.g. by summing them. If sentences (4) and (5) *had* occurred with 'cat' the vector for 'cat' would have been (1 1 0 2). To keep the example simple, however, we are assuming that each word occurred exactly once, so the final vector representations for cat', dog' and banana' are given by (4'), (5') and (6').

Vector representations make the inverted test feasible because they partition all possible sentences into equivalence classes. For example, (4') also represents

We had supper and gave the cat milk (4a)
Our cats like milk, but we prefer steak (4b)
We drank all the milk and gave the cat peanuts (4c)

in addition to the potential infinity of other sentences guaranteed by the productivity of syntax. The fact that (4) and (4a) have equivalent vectors illustrates that the representation is invariant to syntactic variation, which is desirable if we wish to reflect content independently of form. Sentence (4b) illustrates the importance of choosing good context words; we shall see that 'we' is not a good choice of context word because it occurs in so many sentences it cannot be informative about the content of any one in particular. Finally (4c) shows that the vector representation inevitably covers a lot of extraneous material.

The co-occurrence vectors are points in a 4-dimensional semantic space. In this space, the fact that 'cat' is more substitutable with, and by hypothesis more semantically similar to, 'dog' than 'banana' is reflected by the relative distances between points; 'banana' is $\sqrt{3}$ away from 'cat' and 'dog' whereas 'cat' and 'dog' are only $\sqrt{2}$ apart. It is also reflected in the angles between vectors; 'banana' is orthogonal to 'dog' and 'cat'. In this particular case it is clear that the dimensions of each vector are not all equally important for determining similarity relations. For example, removing any of columns 1, 2 and 3 from each vector will preserve distance ranking and orthogonality. Only a subspace is strictly necessary.

Although deliberately small, this example shows the essential components of semantic space models. A semantic space requires a set of context words to serve as a basis ('we', 'clean', 'eat' and 'milk'), a measure of association to represent words occurring in surrounding sentences, a distance or similarity function (Euclidean distance or angle). We might also require a transformation (shrinking the dimensionality of the space by removing one or more of the vector elements).

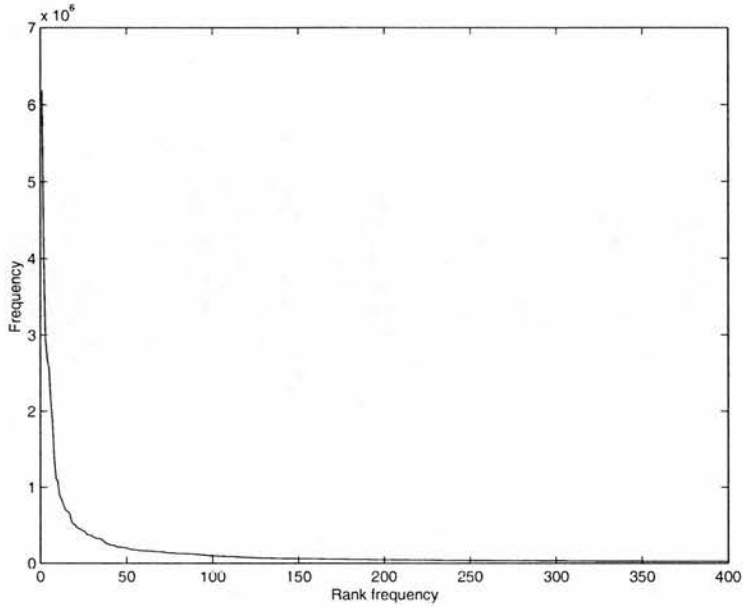

Figure 4.1: Occurrence frequency plotted against rank in a frequency list for the 400 most frequent lemmas in the BNC.

The next section presents a more detailed account of semantic space construction and deals with the problems that arise from scaling up this simple example to more realistic problem sizes where Zipf's law generates significant problems for a straightforward computational implementation of the inverted test.

4.2 Theoretical Foundations

Zipf's law

Zipf's law (Zipf, 1949; Mandelbrot, 1954; Li, 1992) states that the number of times a word occurs is proportional to the reciprocal of its rank frequency. Figure 4.1 shows the empirical frequency distribution for the 400 most frequent lemmas in the British National Corpus, a 100 million word corpus of British English. The 10 most frequent lemmas in the BNC are 'the', 'be', 'of', 'and', 'to', 'a', 'in', 'have', 'that' and 'it'. They constitute slightly over one quarter of all tokens in the corpus[3]. In general the most

[3] 25974687 / 99985962 $\approx$ 0.26

frequent words of English are grammatical functors or closed class words (Cann, 1996).

The power scaling of Zipf's law ensures that the vast majority of words occur very infrequently, creating a severe sparse data problem for statistical language models. In addition even when sufficient data are available, statistical techniques that assume Normal or near Normal distributions are often inappropriate given widely differing word frequencies. Thus, despite the intuitive simplicity of the inverted replacement test, semantic spaces require relatively complex and novel mathematical machinery to deal with the effects of Zipfian distributions. The rest of this chapter is devoted to developing that machinery.

4.2.1 The Theory of Semantic Spaces

A semantic space model is method of assigning each word in a language to a point in a real finite dimensional vector space (Halmos, 1987). Formally it is a quadruple $\langle \mathtt{A}, \mathtt{B}, \mathtt{S}, \mathtt{M} \rangle$: $\mathtt{B}$ is a set $b_{1 \ldots D}$ of basis elements that determine the dimensionality D of the space and the interpretation of each dimension. $\mathtt{B}$ is often a set of words, though in this work we use word stems, or lemmas. Other researchers have used a variety of larger linguistic units from paragraphs to whole documents. $\mathtt{A}$ specifies the functional form of the mapping from co-occurrence frequencies between particular basis elements and each word in the language so that each word is represented by a vector $\mathbf{v} = [\mathtt{A}(b_1, t), \mathtt{A}(b_2, t), \ldots, \mathtt{A}(b_D, t)]$. $\mathtt{S}$ is a similarity measure that maps pairs of vectors onto $\mathcal{R}^1$ to represent the similarity or distance between them. $\mathtt{M}$, is a transformation that takes one semantic space and maps it onto another, for example by reducing its dimensionality. Various choices for these elements are possible, and lead to rather different spaces. In the following sections we consider the implications of different choices of $\mathtt{A}, \mathtt{B}, \mathtt{S}$ and $\mathtt{M}$.

4.2.2 $\mathtt{A}$: Lexical Association Function

Zipf's law suggest that using vectors of co-occurrence counts directly may not be a good choice when constructing a semantic space. To see why, consider two words t_1 and b with occurrence probabilities $p(t_1)$ and $p(b)$ in a corpus of N words. If t_1 and b have *no* semantic relation to each other, then they will be distributionally related to one another only through their syntactic properties e.g. by the fact that they are both nouns. For simplicity can we ignore any residual syntactic dependence and model their empirical

frequencies $f(t_1)$ and $f(b)$ as independent binomially distributed random variables

$$\begin{aligned} f(t_1) &\sim \mathcal{B}(p(t_1), N) \\ f(b) &\sim \mathcal{B}(p(b), N). \end{aligned}$$

In this idealisation t_1 and b are perfectly distributionally independent; the expected frequency of the bigram $\langle b_1, t \rangle$ is $N\,p(b, t_1) = N\,p(t_1)p(b)$. This expected count is linear in the frequency of t_1, so for a fixed basis of context words, when $f(t_1) \ll f(t_2)$ then the absolute values of the t_1 and t_2 vectors will be quite different. However, when $f(t_1) \approx f(t_2)$ then the vectors will be very similar. This means that the distance between t_1 and t_2 in semantic space depends not on tending to co-occur with the same sets of representative words (since they occur with everything completely at random), but on how similar in frequency they are.

The upshot for models such as HAL that use vectors of counts that are not corrected for chance is that distances will have a frequency bias. That is, proximity on semantic space will be partly due to distributional similarity, and partly due to relative frequency. Since it is unlikely that semantic similarity depends on relative frequency, we need a lexical association function $\mathtt{A}(t, b)$ to map raw co-occurrence frequencies onto a less biased measure of association.

The LSA model uses an entropy-based function

$$\begin{aligned} \mathtt{A}(b, t) &= \frac{\log(f(b, t) + 1)}{-\sum_{i=1}^{d} \lambda \log(\lambda)} \\ \lambda &= f(b \mid i) / \sum_{j=1}^{d} f(b \mid j) \end{aligned} \tag{4.1}$$

where $f(b \mid i)$ is the frequency of b in document i and $f(b, t)$ is the number of times t occurs in the same document as b. λ converts counts of the occurrence of b over a document collection into a probability distribution. (In a balanced corpus such as the BNC documents are analogous to changes of author and genre.) The denominator in $\mathtt{A}(t, b)$ is the entropy of this distribution; the entropy is maximised when b is equally likely to occur in every document and minimised when it occurs only in one. The numerator is a logged co-occurrence count, with 1 added to guarantee positivity. Logging $f(b, t)$ emphasises differences between small counts and damps large counts from words with high rank frequencies.

If b is a grammatical functor, e.g. 'we', then it will occur frequently and uniformly across documents, leading to a large entropy term. The entropy term will shrink $\mathtt{A}(t, b)$,

	Target	Non-target
Context	$f^W(b, t)$	$f^W(b, \neg t)$
Non-context	$f^W(\neg b, t)$	$f^W(\neg b, \neg t)$

Table 4.1: Co-occurrence frequency within a window of target, context and all other words. $\neg t$ represents a word that is not t.

reducing its influence on any subsequent similarity measure. In contrast, a technical term or content word may occur infrequently and only in specific documents. It will have a low entropy and its co-occurrence count will be less affected by the log.

LSA's lexical association function is designed to allow many context elements into the distance calculation. However, only informative elements significantly affect the calculation. The measure works well in practice, but its probabilistic foundations are unclear; entropy weighting has clear information-theoretic justification, but logging an augmented co-occurrence count is less obviously motivated. Also, chance co-occurrence is not taken into account, except perhaps that large counts are shrunk proportionally more by the log.

A new lexical association function

In order to take into account chance co-occurrences to create $\mathtt{A}(b, t)$ we must first be able to estimate them. The relevant frequencies can be summarised in a contingency table (Table 4.1) where all frequencies are computed over a W word window ($W/2$ words either side of the target word). $\neg t$ represents any word that is not t, $\neg b$ represents a word that is not the context word and $f(\neg b, t)$ is the number of times a word that is not the context word occurs among the $W/2$ words either side of the target. $f^W(b, t)$ is the regular co-occurrence count with the superscript W marking the fact that, unlike $f(b)$ and $f(t)$, the frequency is computed over a window of size W.

Computing the cell counts is straightforward because they are all functions only of

$f^W(b,t)$ itself, the occurrence frequencies of t and b, the window size W, and the corpus length N:

$$\begin{aligned} f^W(b,\neg t) &= Wf(b) - f^W(b,t) \\ f^W(\neg b,t) &= Wf(t) - f^W(b,t) \\ f^W(\neg b,\neg t) &= WN - (f(b,\neg t) + f(\neg b,t) + f^W(b,t)). \end{aligned} \tag{4.2}$$

To prove these expressions we first define the *distance bigram* frequency $d_i(b,t)$, which is the number of times b occurs exactly i words before t. It is important to note that unlike f, the order of t and b in text matters when computing d_i because typically $d_i(b,t) \neq d_i(t,b)$. Normal bigram frequencies are special cases of distance bigrams where $i = 1$.

We then consider a $W \times 2 \times 2$ table, that embeds W subtables, one for each position b can take with respect to t within the window. Subtables where $i = 1 \ldots W/2$ contain $d_i(b,t)$, $d_i(\neg b,t)$, $d_i(b,\neg t)$ and $d_i(\neg b,\neg t)$ and subtables where $i = W/2+1 \ldots W$ contain $d_i(t,b)$, $d_i(\neg t,b)$, $d_i(t,\neg b)$ and $d_i(\neg t,\neg b)$. It is not necessary to explicitly compute co-occurrence frequencies for non-targets and non-context words individually[4]. For example

$$d_i(\neg b,t) = f(t) - d_i(b,t)$$

because $f(t) = d_i(\neg b,t) + d_i(b,t)$ for any i and any b. Similarly

$$d_i(\neg b,\neg t) = N - (d_i(b,t) + [N - d_i(\neg b,t)] + [N - d_i(b,\neg t)]).$$

Lastly note that

$$f(b,t) = \sum_{i=1}^{W/2} d_i(b,t) + d_i(t,b). \tag{4.3}$$

Since the position variable indexed by i is not considered relevant to measuring substitutability we need to marginalise. This amounts to collapsing the three-way table over the position dimension. Equation 4.3 shows that $f^W(b,t)$ is effectively already

[4]Strictly this is true when $i = 1$ or when no context word occurs on both sides of target word within the window. Excluding very high frequency words from the context set reduces the probability of this occurring, as do natural expressive constraints on using the same word multiple times in quick succession in normal language, so the probability of error using this method may be expected to be negligible.

collapsed. The other two cell types are then

$$
\begin{aligned}
f(\neg b, t) &= \sum_{i=1}^{W/2} d_i(\neg b, t) + d_i(\neg t, b) && (4.4)\\
&= \sum_{i=1}^{W/2} (f(t) - d_i(b, t) + f(t) - d_i(t, b)) && (4.5)\\
&= \sum_{i=1}^{W/2} (f(t) - d_i(b, t)) + \sum_{i=1}^{W/2} (f(t) - d_i(t, b)) && (4.6)\\
&= (W/2) f(t) - \sum_{i=1}^{W/2} d_i(b, t) + (W/2) f(t) \sum_{i=1}^{W/2} d_i(t, b) && (4.7)\\
&= W\, f(t) - \sum_{i=1}^{W/2} (d_i(b, t) + d_i(t, b)) && (4.8)\\
&= W\, f(t) - f^W(b, t) && (4.9)
\end{aligned}
$$

and

$$
\begin{aligned}
f(\neg b, \neg t) &= \sum_{i=1}^{W/2} d_i(\neg b, \neg t) + d_i(\neg t, \neg b) && (4.10)\\
&= \sum_{i=1}^{W/2} d_i(\neg b, \neg t) + \sum_{i=1}^{W/2} d_i(\neg t, \neg b) && (4.11)\\
&= (W/2)N - \left(\sum_{i=1}^{W/2} d_i(b, t) + d_i(\neg b, t) + d_i(b, \neg t) \right) + \\
&\quad (W/2)N - \left(\sum_{i=1}^{W/2} d_i(t, b) + d_i(\neg t, b) + d_i(t, \neg b) \right) && (4.12)\\
&= W\, N - \left(\sum_{i=1}^{W/2} d_i(b, t) + \left[W/2 f(t) - \sum_{i=1}^{W/2} d_i(b, t) \right] + \left[W/2 f(b) - \sum_{i=1}^{W/2} d_i(b, t) \right] + \right. \\
&\quad \left. \sum_{i=1}^{W/2} d_i(t, b) + \left[W/2 f(b) - \sum_{i=1}^{W/2} d_i(t, b) \right] + \left[W/2 f(t) - \sum_{i=1}^{W/2} d_i(t, b) \right] \right) && (4.13)\\
&= W\, N - (f^W(b, t) + [W\, f(t) - f^W(b, t)] + [W\, f(b) - f^W(b, t)]) && (4.14)
\end{aligned}
$$

From these results about frequency counts, obtaining the corresponding probabilities (Table 4.2) is straightforward: the count in each cell is divided by WN. The proof of this fact is similar to those presented above but not more illuminating.

	Target	Non-target
Context	$p^W(b,t)$	$p^W(b,\neg t)$
Non-context	$p^W(\neg b,t)$	$p^W(\neg b,\neg t)$

Table 4.2: Co-occurrence probability for context word b and target b.

From Table 4.2 we can see that the odds of seeing t rather than some other word when b is present are $p^W(b,t)/p^W(b,\neg t)$, and the odds of seeing t in the absence of b is $p^W(\neg b,t)/p^W(\neg b,\neg t)$. Therefore if the presence of b *increases* the probability of seeing t then the odds ratio (Agresti, 1996)

$$\begin{aligned}\theta(b,t) &= \frac{p^W(b,t)/p^W(b,\neg t)}{p^W(\neg b,t)/p^W(\neg b,\neg t)} \\ &= \frac{p^W(b,t)\,p^W(\neg b,\neg t)}{p^W(b,\neg t)\,p^W(\neg b,t)}\end{aligned} \tag{4.15}$$

is greater than 1. So if 'cat' and 'milk' are positively associated then $\theta(\text{milk, cat}) > 1$ because the presence of 'cat' increases the chances of seeing 'milk' in the window. When the presence of b makes no difference to the probability of seeing t then $\theta = 1$ and we can conclude that b and t are distributionally independent. We might expect that $\theta(\text{we, cat}) = 1$. Finally, if $\theta < 1$ the presence of t makes seeing b less probable.
We can estimate the odds ratio from Table 4.1:

$$\hat{\theta}(b,t) = \frac{f^W(b,t)\,f^W(\neg b,\neg t)}{f^W(b,\neg t)\,f^W(\neg b,t)}.$$

Since it is a ratio $\hat{\theta}$ may increase infinitely in the positive direction but is bounded below by 0. Logging makes $\hat{\theta}$ symmetrical around 0 and asymptotically Normal. This is particularly useful for analyses of variance where approximate normality is desirable. We show how to use ANOVA methods to choose context words below. Normality is particularly difficult to achieve when using raw co-occurrence frequency because of the

skew induced by Zipf's law. Box and Tiao, 1973 point out that variance comparisons are particularly sensitive to non-Normality[5].

We can interpret the magnitude of $\log \hat{\theta}(b,t)$ directly as a measure of associative strength between t and b. The sign gives direction to the association, although we are really only interested in words that occur more often than chance around t because they are the words that have the most informative distributional profile. The *most* informative words for t are those that occur only in its context, e.g. t='sealed' and b='hermetically'. Instances of word pairs like 'hermetically' and 'sealed' are concordances, or collocations, and are of considerable interest to lexicographers (See Manning and Schütze, 1999, chapter 5 for a review). Consequently, the log odds ratio also provides a method of finding collocations between words. Previous work has used pointwise mutual information, log-likelihood ratios, and T-tests. Since by symmetry these alternative measures can also be lexical association functions, and because previous work used the log-likelihood ratio (McDonald and Lowe, 1998), we review them briefly below.

Pointwise Mutual Information

The pointwise mutual information $I(b,t)$ between t and b is

$$I(b,t) = \log \frac{p^W(b,t)}{Wp(b)p(t)} \tag{4.16}$$

and can be estimated using the frequencies in Table 4.1. $I(b,t)$ measures how much information an occurrence of b contains about t (and vice versa since it is symmetric). If b occurs with t no more often than would be expected by chance then $p^W(b,t) = Wp(b)p(t)$ and $I(b,t) = 0$, so the mutual information measure effectively factors out random co-occurrences. However, if t and b always occur together then $p^W(b,t) = p(b)$ and $I(b,t) = \log 1/p(t)$, so the less frequent b and t are the larger their association is. In contrast, changing the marginal probabilities of t or b is equivalent to adding a constant value to rows or columns of the contingency tables above (Bishop et al., 1975). It is straightforward to confirm that this change makes no difference to the value of $\hat{\theta}$.

[5]Critical regions for equality of variance overlap those for kurtosis, p.203.

The Log-likelihood Ratio

In classical statistics we can judge the relative plausibility of two models by examining the ratio of their maximised likelihood functions (Agresti, 1990)

$$\lambda = \frac{\max L(\Theta_{M1}; \text{data})}{\max L(\Theta_{M2}; \text{data})}$$

where Θ_{M1} denotes the parameters of model M1 and $L(\Theta_{M1}; \text{data}) \propto p(\text{data} \mid \Theta_{M1})$. The quantity $-2 \log \lambda$ is then asymptotically χ^2-distributed, so we can use a significance test to see whether M1 is more probable than M2. Alternatively λ itself can be interpreted directly as a measure of how much more plausible M1 is than M2.

Dunning, 1993 compared two models of the relation between t and b

M1: (association) $p^W(b \mid t) \neq p^W(b \mid \neg t)$

M2: (no association) $p^W(b \mid t) = p^W(b \mid \neg t)$.

If t and b are associated then b is distributed differently depending on whether t is nearby or not. Consequently, $M1$ has two parameters, $p^W(b \mid t)$ and $p^W(b \mid \neg t)$ to cover either situation. $M2$ on the other hand only requires one, $p(b)$ because b behaves the same way irrespective of t. Dunning sets all three parameters to their maximum likelihood values and uses $\log \lambda$ to measure the strength of association between t and b.

The measure takes chance into account because it implicitly compares the observed co-occurrence frequencies with the co-occurrence frequencies that would be expected by chance. For example, the expected value of the top left cell in Table 4.1 is $Wf(t)f(b)/N$ under $M1$ but $f^W(b,t)$ under $M1$. The more such cell estimates differ between models the larger $\log \lambda$ becomes. In fact, the log-likelihood ratio is the standard model selection criterion for comparing different hierarchical log-linear models. Table 4.2.2 has only two possible models: In standard notation the models are [PT] and [P][T], corresponding exactly to M1 and M2 above.

$\log \lambda$ has been used with some success as a measure a lexical association (Dunning, 1993; McDonald and Lowe, 1998). Empirically it seems that using log-likelihood ratios as vector elements generates very similar results to using log odds-ratios (Compare the analysis of Moss *et al.*'s data in the results chapter to that performed by McDonald and Lowe, 1998). This is to be expected since both measures take chance co-occurrences into account.

Using the log odds-ratio

For the simulations below we discard all negative values of $\hat{\theta}$. There are two reasons for this choice.

1. We assume that 'positive' associations are more psychologically salient.

2. The second reason is more speculative. That b_1 occurs with t more often than chance is more salient and more likely to be represented than the fact that b_2 occurs with t *less* than would be expected by chance, simply because every time b_1 occurs in a window with t then the unrelated b_2 word cannot. This forces $\log \theta(b_1, t)$ below 0, and has a similar effect on the ratios of all other unrelated words. A large number of log odds-ratios can therefore be expected to be less than 0 when a few are much above it. Consequently the negative values can be discarded.

The second reason is more speculative because it presupposes fixed word occurrence probabilities. The argument goes through if b 'needs' to occur $f(b)$ times, and must find suitable locations in the text to do so. Perhaps a more intuitive model of text generation is that when t has a number of strong associates then $f(b)$ simply drops because b appears all the places it would usually appear but seldom in places that would bring it near t.

It is worth noting that vectors of odds-ratios do seem to exhibit the pattern of results that are predicted by reason 2. In Moss *et al.*'s materials which were balanced for frequency, the number of zeros in each vector correlated with the average value of the remaining log-odds ratios, r=0.459 p<.001. The correlation dropped but remained reasonably sized when four extreme log odds-ratios values were trimmed, r=0.223 p<.001, so this provides some support for reason 2.

4.2.3 `B` : Choosing a Basis

When choosing basis elements for a semantic space there is a tradeoff between choosing words that are representative of sentence content, but may not give reliable count statistics due to their low frequency, and choosing high frequency words that provide reliable statistics but appear in almost every sentence of the language. The tradeoff is an instance of the bias-variance dilemma in statistical learning theory (Geman et al., 1992).

Bias and Variance

Every statistical model is able to represent a subset of the class of possible hypotheses about data. The range of hypotheses is typically controlled by the model's structure and by a set of adjustable parameters. More flexible models can represent more hypotheses and are said to have less *bias*. In contrast, a very flexible model will require a large amount of data to pin down values for its parameters. When there is not enough data compared to the number of parameters, parameter estimates may be optimal for the particular data set the model was trained on, but will fail to generalise to new data. A model that 'overfits' in this way is said to have high *variance*. Model variance can be decreased at the cost of adding bias e.g. by constraining or removing parameters. Bias can be decreased by making the model more flexible, at the cost of needing more data to cope with increased variance.

In a semantic space the vector elements, $\mathtt{A}(b,t)$ are parameters that estimate the amount of association between b and t on the basis of observed data $f^W(b,t)$ (and for our model also $f(t)$ and $f(b)$). When choosing the basis elements $b_1 \ldots b_D$ to count, we can define a highly biased model by choosing only very high frequency words. Co-occurrence counts for high frequency words are very reliable because high frequency words appear in nearly all sentences. This biased model will have very low variance since each $\mathtt{A}(b,t)$ is a well-determined parameter because $f^W(b,t)$ is large enough to provide a reliable estimate of $p^W(b,t)$. However, every vector will be similar because all words in the language tend to occur with the high frequency words in the basis, irrespective of their distributional profile. Consequently, distances between words will be extremely similar and vectors in the biased model will fail to reflect important distributional differences (Imagine trying to distinguish 'cat' and 'banana' using context words 'a', 'the' and 'it').

Alternatively, if low frequency content words are chosen as basis elements then vectors will be more highly informative and distances in the space will be able to reflect subtle distributional similarities. Unfortunately this model has high variance because the co-occurrence counts needed to determine $\mathtt{A}(b,t)$ are unreliable. For small values of $p(b)$, whether $f^W(b,t)$=5, 10 or 0 can vary depending on the corpus at hand. This is not helpful when the aim is to make general statements about the relation between two words in English, rather than just in the BNC. Variance can always be decreased by providing more data, but Zipf's law suggests a power relation between the

amount of new text that would need to be found and the reduction in co-occurrence count variability.

In practice, bias and variance are not equally problematic when building semantic space models. If very low frequency or otherwise unreliable basis elements are avoided then distances between vectors will be representative. On the other hand, as long as the lexical association function takes into account chance occurrences, it will not significantly harm the model to add high frequency basis elements. Although they are noise in the model and we avoid them if possible, high frequency elements should not make much difference to distance calculations because their similar values contribute only small terms to similarity measures. Including large numbers of possible words and relying on the lexical association function to down-weight less useful elements is the approach taken in LSA (see above).

Latent Semantic Analysis

LSA overcomes the problem of low frequency basis elements by choosing paragraphs or articles as elements of B, and by weighting co-occurrence counts according to their expected informativeness. Choosing larger textual units as basis elements reflects LSA's origins in document retrieval, where it is called Latent Semantic Indexing. In vector space document retrieval (Salton and McGill, 1983) the problem is to retrieve all and only the documents in a collection that are semantically related to the words in a user's query. Entire documents are represented in terms of word frequency counts over elements of B. Although counts over a document are typically more reliable than co-occurrence statistics in a window, Zipf's law still leads to the problem of choosing appropriate words.

In document retrieval the problem is expressed in terms of precision and recall. For each query vector we assume that there is a particular set of documents that should be returned. We then set a threshold for the similarity measure, say a cosine of 0.7, and return all documents that have vectors with cosines of more than 0.7 with the query. Then the *Precision* is the proportion of relevant versus irrelevant documents returned as the result of a query. *Recall* is the proportion of the set of relevant documents that are actually returned. If several documents are relevant, then returning just one relevant document shows exemplary precision but poor recall, and returning all the documents shows perfect recall but no precision. For a fixed similarity measure S, the precision

and recall of a retrieval system depends entirely on the choice of basis elements.

Very frequent words tend to occur with regular spacing throughout any corpus and provide a relatively reliable guide to the distributional profile of any particular word, but since they will not be semantically specific, they will not serve to pick out any particular set of documents. Consequently, they will help recall but hinder precision. But as words become less frequent they tend to become increasingly informative about their contexts. Relatively infrequent words e.g. technical terms, tend to be semantically specific and occur in a very restricted range of contexts. The presence of the word thus becomes a very good indicator of that context and when it is used as a context word it effectively distinguishes targets that can occur in that context from those that cannot. But co-occurrence counts for infrequent words are unreliable and lead to high variance estimates of lexical association and distributional similarity. In other words such words will help precision by being very specific, but hinder recall by being difficult to estimate and irregularly distributed across the corpus.

Precision and recall are alternative expressions of the bias/variance tradeoff. This is because high frequency basis elements generate document vectors that are very close together in semantic space. The result is a model with good recall and terrible precision because many irrelevant documents will have cosines with the query that are above threshold. Alternatively, low frequency basis elements create a model with good, though erratic, precision because individual content words can be very informative about document similarity but terrible recall because the vectors are widely dispersed in the space due to unreliable counts. Just as we might ask for a semantic space with as little bias as possible for reasonable variance, information retrieval researchers require vector space models with as much precision as possible for a fixed amount of recall.

When LSA is used as a psychological model Landauer *et al.* reformulate the document retrieval problem as one of retrieving a set of semantically similar words from a space defined by documents. The original document retrieval problem gives a simple answer to the question of how to choose basis elements: Basis elements are documents, so choose the ones that you have. When the task is inverted the documents simply become basis elements. In subsequent work Landauer *et al.* have used smaller basis elements such as paragraphs.

In this context LSA may be more biased than a window-based approach. If substitutability really is a reliable indicator of semantic similarity, then many of the word

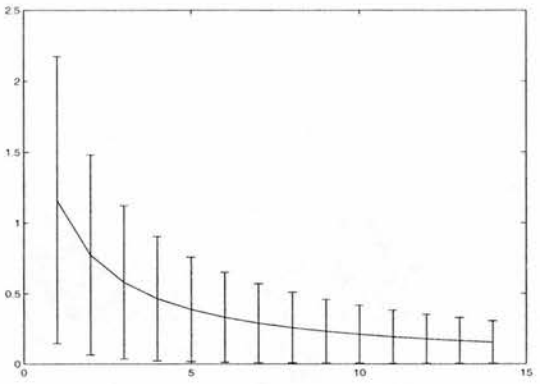

Figure 4.2: An example of Burgess *et al.*'s column sum method for choosing basis words. Expected column means based on expected co-occurrence counts between each of 14 hypothetical unrelated words. To estimate means and variances for a corpus of N words, multiply all quantities by N. Error bars represent expected column variances. Word counts are assumed to be independent and Binomially distributed with occurrence probabilities in accordance with Zipf's law, ranging from 0.5 to 0.0667. In Burgess *et al.*'s method, words with the largest column variances (here error bars) are chosen as basis elements.

counts in each document are simply noise and the detailed distributional information contained in the semantic replacement test will be obscured by counts from unrelated parts of the document.

The Hyperspace Analogue to Language

In HAL, elements of B are chosen by compiling a 70,000×70,000 matrix of word co-occurrences and discarding the columns of lowest variance[6]. Consistent with Zipf's law, column variance decreases sharply with the frequency of the word corresponding to the column (Lund et al., 1995). For each set of experimental stimuli, Burgess *et al.* compute variances over each vector element and retain only the most variant.

This method is difficult to analyse, particularly because the basis is recomputed for each experiment. To the extent that it is analytically tractable, we can show that it has a frequency bias. However, a more pressing methodological question is why words with maximally variant co-occurrence counts should have properties desirable for a semantic

space. The following section explores the nature of the frequency bias, and suggests a variant on Burgess *et al.*'s method that takes frequency into account.

If b and t are unrelated then we can model them as Binomially distributed (see above). For simplicity we set $W = 1$. The variance of the frequency count under independence is then

$$\begin{aligned} \text{Var}\, f^W(b,t) &= Np(t)p(b)(1 - p(t)p(b)) \\ &= Np(t)p(b) - Np(t)^2p(b)^2. \end{aligned}$$

The expected variance of $f^W(b,t)$ increases quadratically in $p(b)$. The expected variance of the elements of a column of such counts is the same as the variance of the column sum i.e. the sum of the individual variances (Feller, 1950). Figure 4.2 shows the expected variances for a 14×14 table of co-occurrence counts for perfectly unrelated words with occurrence probabilities ranging from 0.5 to 0.0667. Even completely unrelated words will show distinct structure in their column variances, but this is entirely due to their baseline frequencies.

There are always two possible causes for a high column variance. The first cause is simple frequency as shown in Figure 4.2. The second reason is that the words are in fact distributionally related. In the generalised linear model literature variance in a dependent variable that is *more* than would be expected from the Binomial definition is taken as a sign of cases 'clumping' (Agresti, 1990), and is often dealt with by adding an additional variance term. Consequently in the linguistic case it is possible to interpret unexpectedly large variance as a sign that the Binomial assumption has failed, and that two words are in fact related (they clump together). This suggests yet another method to take chance into account when measuring associations. Perhaps, choosing columns according to the amount of extra variance would be a good way to choose basis elements that are related to the words in each experiment. This scheme would be one charitable interpretation of the Burgess *et al.* method. However this scheme is not that method. Burgess *et al.*'s method uses column variance alone to decide on basis elements. For a word that is distributionally related to some of the experimental materials to make it into the final lineup it must be strongly associated enough that its observed column variance moves it into the window of very high variance words at the upper end of the frequency table. In other words, it is not enough to be twice as

[6] Co-occurrences are also weighted by distance, but this does not affect the following argument.

variant as would be expected by chance, a word must be as many times more variant as it takes to have a variance that is absolutely high; lower frequency words have to work harder and unrelated but high frequency words will get chosen anyway.

This analysis of Burgess *et al.*'s methods predicts that, in the absence of strong association, the variance of a column corresponding to some candidate element will correlate strongly with that element's frequency. This was tested by taking candidate basis elements of rank 100 to 600 in the frequency list for the BNC, and experimental stimuli from McKoon and Ratcliff's graded priming study (discussed in detail in the next chapter). We expect that the levels of actual association (corrected for frequency) between these candidates and the experimental stimuli to be low because the words are so frequent that they provide little information about context. Indeed the lower frequency candidates tended to be more associated; the variances of columns of *log odds-ratios* were negatively correlated with candidate frequencies, r=-.317 $p<.001$, with column means closer to 0 as frequency increased. In contrast candidate frequencies correlated positively with column variance for co-occurrence counts, r=.8553 $p<.001$. This number gives a quantitative estimate of the amount of frequency bias in the method.

If all candidates were completely unassociated with the experimental materials r would be 1. However there were 22 distinct outliers that reduced the correlation coefficient. We removed them and looked at the column variance of the corresponding log odds-ratios in the expectation that these words were associated enough to be visible above the chance co-occurrence count levels. If the outliers were also significantly associated according to a frequency-corrected measure, then the method is partly vindicated since these words are more likely to be chosen. Unfortunately the outlying words were not particularly strongly associated to the experimental stimuli; the mean column log odds-ratio across the outliers was 0.08214 (slightly negatively associated) with standard deviation 0.2067, and there was no systematic relation in variance structure among the log odds-ratios. What makes these words outliers is presently unclear, but it is not high levels of association with the experimental materials. In conclusion, even the outlying candidates that have particularly large column variances and are destined to be chosen above their higher frequency neighbours in Burgess *et al.*'s method are not particularly more associated with the experimental materials than any other candidates, and we have no further understanding of why the method produces reasonable words.

A new method for choosing a basis

When a lexical association measure takes chance occurrence into account there is no theoretical reason to constrain the dimensionality of the semantic space. Two words that have no association will take on a log odds-ratio close to 0, so that particular word combination will have very small if any effect on subsequent distance or angle combinations. This approach is taken when constructing LSA models; the entropy term downweights the estimates of association for uninformative word combinations (see above). However, although estimates of log odds-ratios tend to 0 in the absence of association, they do so asymptotically. When counts are very low they are still unreliable. Consequently there are advantages in practise to restricting basis elements to those that are *reliable*.

To quantify reliability we treat basis elements like human raters and use standard rater reliability ANOVA models to assess their reliability[7]. To find reliable context words we first choose several thousand candidate basis elements from the high frequency portion of the corpus, excluding stop words. We then pick randomly another set of words called dummy targets. Using the log odds-ratio, we create vectors for each dummy target using the candidate context words, over k disjoint sections of the corpus. In these experiments we used k=4 corpus sections containing 10M words each from the first half of the BNC.

Reliable candidates will generate k similar vectors corresponding to each dummy target, whereas unreliable candidates will behave differently depending on which section of the corpus counts are taken from. Since we are looking for low variance choices of basis elements, we can use the fact that high variance choices will be sensitive to the 'training data' (the different corpus sections) to screen out inappropriate words. We use a within 'subjects' (the candidates) ANOVA to test whether each candidate generates significant variation in vector elements between the k tests. Context words for which we cannot reject the null hypothesis of no variation between corpus sections are retained. With a rather conservative critical significance level of 0.1, the procedure generates 536 context words with mean frequency 219.367 and median frequency 159.77 per million words. Basis elements are listed in Appendix A.

In fact the choice of basis elements appears to be the least important choice in constructing a semantic space. The work presented here has been repeated with several

[7]The methods described here are a development of those used in McDonald and Lowe, 1998.

subsets of the chosen words and generated essentially identical substantive results.

4.2.4 S : Similarity Measure

There are essentially two choices for comparing vectors in semantic space: Euclidean distance and the cosine measure. LSA uses the cosine, and HAL uses Euclidean distance on normalized co-occurrence vectors. The experiments described in this thesis all use the cosine measure.

For two vectors $\mathbf{v}$ and $\mathbf{w}$ in a D-dimensional basis, the squared Euclidean distance $\| \mathbf{v} - \mathbf{w} \|^2$ is simply related to the cosine $\rho_{\mathbf{vw}}$ of the angle between them,

$$\| \mathbf{v} - \mathbf{w} \|^2 = \sum_{i=1}^{D} (v_i - w_i)^2 \tag{4.17}$$

$$= \| \mathbf{v} \|^2 + \| \mathbf{w} \|^2 - 2 \frac{\mathbf{vw}}{\| \mathbf{v} \| \| \mathbf{w}} \tag{4.18}$$

$$= \| \mathbf{v} \|^2 + \| \mathbf{w} \|^2 - 2 \rho_{\mathbf{vw}} \tag{4.19}$$

where $\| \mathbf{w} \|^2 = \sum_i^D w_i^2$ is a vector length. From this equation it can be seen that $\| \mathbf{vw} \|^2 \propto \rho_{\mathbf{vw}}$ only when $\mathbf{v}$ and $\mathbf{w}$ are standardised in length.

One advantage of the cosine is that it ranges between -1 and 1, and so removes any arbitrary scaling induced by the range of A and the number of elements in B. When A is simple co-occurrence the cosine is also less sensitive than Euclidean distance to extreme values induced by widely differing basis element frequencies, although a good choice of A should avoid this problem.

4.2.5 M : Model

A semantic space is fully functional when a B, A and S have been specified. However, it is possible to build a more structured mathematical or statistical model. In LSA the model consists of a projecting vectors into a linear subspace of B using singular value decomposition. Landauer *et al.* note that this greatly improves the model's fit to data. Below we present an analysis of LSA in terms of the variance structure in the lexical association data, and review some recent probabilistic extensions. In the next chapter we present a new model for capturing variance structure using non-linear projections into a low-dimensional subspace.

LSA

LSA decomposes a matrix of lexical association values $\mathbf{A}$ using singular value decomposition (Golub and Van Loan, 1989)

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\mathsf{T} \tag{4.20}$$

$$= \sum_{i=1}^{r} \sigma_i \mathbf{u}_{(i)} \mathbf{v}_{(i)}^\mathsf{T} \tag{4.21}$$

where $\mathbf{u}_{(i)}$ denotes the ith column of $\mathbf{U}$. Note that SVD is symmetrical because $\mathbf{A}^\mathsf{T} = \mathbf{V}\Sigma^\mathsf{T}\mathbf{U}^\mathsf{T}$. Treating column and row vectors symmetrically is necessary for term manipulation in document retrieval applications and distinguished LSA from most other statistical methods for dimension reduction. Typically, rows of the matrix are data points and positions in column space are not meaningful.

$\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$ and $\Sigma = \text{diag}(\sigma_1 \ldots \sigma_r)$ is a diagonal matrix containing the r singular values. Equation 4.21 shows that SVD decomposes $\mathbf{A}$ into a sum of rank 1 matrices, so the number of non-zero singular values gives the rank of $\mathbf{A}$.

Latent Semantic Analysis assumes that the elements of $\mathbf{A}$ are measurements from a noisy process. Some of the rank 1 matrices in Equation 4.21 then reflect unsystematic variation due to measurement error and the intrinsic rank of $\mathbf{A}$ may be less than r. The relation between intrinsic dimensionality and intrinsic rank is important for understanding LSA, and is discussed further below. To remove unsystematic variation, LSA reconstructs $\mathbf{A}$ using *thin* singular value decomposition. In thin SVD $\mathbf{A}$ is decomposed as in Equation 4.21 and all but the k largest singular values are removed leaving $\Sigma_{[k]} = \text{diag}(\sigma_1 \ldots \sigma_k)$. The matrix is reconstructed

$$\hat{\mathbf{A}} = \mathbf{U}\Sigma_{[k]}\mathbf{V}^\mathsf{T} \tag{4.22}$$

The reconstructed matrix $\mathbf{A}_{[k]}$ is the rank k matrix that is closest to $\mathbf{A}$ in a least-squares sense (Golub and Van Loan, 1989). The reconstruction is optimal, though computationally intensive. Since $\Sigma_{[k]}$ now has k diagonal elements, only k of the columns of $\mathbf{U}$ and $\mathbf{V}$ will affect the reconstruction. The choice of k is empirical and implicitly reflects assumptions about the nature of the process that generates the matrix and the amount of noise that accompanies the process.

It should be noted, however, that talk of noise and measurement error in SVD is only analogical; SVD is a purely algebraic manipulation that has no associated statistical

model.

Intrinsic rank and dimensionality

To understand the relation between intrinsic rank and intrinsic dimensionality, it is useful to consider SVD's connection to principal component analysis which in turn enables a more straightforward connection to statistical models.

The sample covariance matrix of a set of N semantic space vectors defined by rows of $\mathbf{A}$ is

$$\mathbf{S} = \frac{1}{N}\mathbf{A}^\mathsf{T}\mathbf{A} - \bar{\mathbf{a}}\bar{\mathbf{a}}^\mathsf{T} \tag{4.23}$$

where $\bar{\mathbf{a}}$ is a $d \times 1$ vector of column means. For ease of exposition we can subtract the mean values from each row of $\mathbf{A}$, since this does not alter the covariance structure. Then $\mathbf{S} = 1/N\,\mathbf{A}^\mathsf{T}\mathbf{A}$. $\mathbf{S}$ describes the variance structure among vectors of co-occurrence statistics as points in the d-dimensional space defined by the elements of B. Since $\mathbf{S}$ is symmetric it has the spectral decomposition

$$\mathbf{S} = \mathbf{W}\Lambda\mathbf{W}^\mathsf{T} \tag{4.24}$$

$$= \sum_{i=1}^{d} \lambda_i \mathbf{w}_{(i)}\mathbf{w}_{(i)}^\mathsf{T} \tag{4.25}$$

Columns of $\mathbf{W}$ are the eigenvectors of $\mathbf{S}$, and $\Lambda = \mathrm{diag}(\lambda_1 \ldots \lambda_d)$ is matrix of eigenvalues such that $\mathbf{S}\mathbf{w}_{(i)} = \lambda_i \mathbf{w}_{(i)}$. The eigenvectors of $\mathbf{S}$ point in the directions of maximum variance in semantic space, subject to the constraint that they are orthogonal to one another; $\mathbf{w}_{(1)}$ points in the direction of maximum variance, $\mathbf{w}_{(2)}$ is the next most variant direction that is orthogonal to $\mathbf{w}_{(1)}$, and so on. The amount of variance in direction i is given by λ_i.

The eigenvectors of $\mathbf{S}$ provide an *alternative* basis to B in which to locate the semantic space vectors in $\mathbf{A}$. Transforming each co-occurrence vector $\mathbf{a}_i$ into the new space makes each vector dimension independent with variance given by the corresponding eigenvalue

$$\mathbf{x} = \mathbf{W}^\mathsf{T}\mathbf{a} \tag{4.26}$$

This transformation is called principal component analysis, and the $\mathbf{x}$s are the principal components. Since this transformation is only a change of basis it can be reversed

straightforwardly

$$\mathbf{a} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{W}^{\mathsf{T}}\mathbf{a} = \mathbf{I}\mathbf{a} = \mathbf{a} \tag{4.27}$$

because $\mathbf{W}$ has orthogonal columns.

If there are only a few directions of significant variance among the co-occurrence vectors, represented by large eigenvalues, then the remaining directions can be assigned to measurement noise. We can define a dimension-reducing transformation by removing columns of $\mathbf{W}$ corresponding to all but the k largest eigenvalues to give $\mathbf{W}_{[k]}$, a $k \times d$ matrix. Vectors can then be mapped into a lower dimensional space

$$\mathbf{x} = \mathbf{W}_{[k]}^{\mathsf{T}}\mathbf{a}. \tag{4.28}$$

This projects a semantic space vector into a subspace that covers the maximum amount of variation, calculated over all the vectors. Subsequent reconstruction proceeds as before

$$\hat{\mathbf{a}} = \mathbf{W}_{[k]}\mathbf{x} \tag{4.29}$$

with the constraint that it cannot be perfect because some variance information has been lost. However $\hat{\mathbf{a}}$ is the optimal reconstruction of $\mathbf{a}$ in a least squares sense (Pearson, 1901, cited in Tipping and Bishop (1997)). Good reconstruction with $k < d$ then suggests that the *intrinsic* dimensionality of the data is k rather than d. Combining the two transformations above

$$\hat{\mathbf{a}} = \mathbf{W}_{[k]}\mathbf{W}_{[k]}^{\mathsf{T}}\mathbf{a} \tag{4.30}$$

shows that the PCA reconstruction projects $\mathbf{a}$ linearly onto a k-dimensional manifold that is embedded in the original space defined by `B`.

Equation 4.29 has the basic structure of a latent variable model reminiscent of Factor Analysis: An unobserved set of k independent sources with variances $\lambda_1 \ldots \lambda_k$ generates observed points $\mathbf{a}$ by mapping them linearly into a higher dimensional space. All that is lacking is a model of the noise that perturbs the model's reconstruction $\hat{\mathbf{a}}$ to give the observed points. Tipping and Bishop's Probabilistic Principal Component Analysis is essentially this model.

Since both principal component analysis and SVD both give optimal linear projections it is possible to use PCA's statistical interpretation to understand SVD. Equa-

tions 4.21 and 4.25 lead to the following equalities

$$\begin{aligned} \mathbf{A}^{\mathsf{T}}\mathbf{A} &= \mathbf{W}\Lambda\mathbf{W}^{\mathsf{T}} && (4.31) \\ &= \mathbf{V}^{\mathsf{T}}\Sigma^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}} && (4.32) \\ &= \mathbf{V}\Sigma^{2}\mathbf{V}^{\mathsf{T}} && (4.33) \end{aligned}$$

Thus $\mathbf{V} = \mathbf{W}$, which gives one third of the SVD a statistical interpretation as the directions of maximal variance in d-space. Since $\mathbf{A}^{\mathsf{T}}\mathbf{A} = n\mathbf{S}$, and multiplying every element of a matrix by n is equivalent to multiplying each of its eigenvalues by n (Mardia et al., 1979), we can see that the singular values of $\mathbf{A}$ are related to the eigenvalues of $\mathbf{S}$ by

$$\sigma_i^2 = \frac{\lambda_i}{n} \quad (4.34)$$

Thus the squared singular values also have interpretations as variances in d-space. Finally, the principal components of $\mathbf{A}$ are $\mathbf{U}\Sigma$, since $\mathbf{U}\Sigma = \mathbf{AV}$ from the definition of singular value decomposition.

A thin SVD reconstruction of a matrix of semantic space vectors first projects each vector into a subspace defined by the principal directions of variance in the data. This is a dimensionality reduction equivalent (assuming subtracted means) to taking the principal components of each vector. The components are then projected back into the original basis B. (Berry et al., 1995) have recently suggested that LSA may work well without the second part of this process. That is, cosines in the *subspace* should be used to measure the similarity between semantic space vectors. As suggested by its similarities to PCA, LSA can then be seen as a linear latent variable model similar to the models discussed in Chapter 2. The GTM model is then a non-linear extension of LSA, where cosines in the latent space are taken to represent semantic similarity.

4.3 Conclusion

This chapter has argued that semantic space models are an implementation of the statistical replacement test for determining ease of substitution in context, and that the replacement test itself derives from a distributional theory of meaning that emphasises language use.

We also developed a general theory of semantic space and showed how different theoretical choices led to HAL and to LSA. We analysed HAL's methods for generating

lexical associations and choosing context words and found significant frequency biases. We then presented an alternative lexical association measure based on the odds-ratio that factors out the effects of chance in word association. We also introduced a new method for choosing context words using words as raters in a rater-reliability framework and showed how the algebraic process of thin Singular Value Decomposition used in LSA relates to statistical characterisations of word vectors in semantic space, and to the latent variable models considered in Chapter 2. In the next chapter we demonstrate the performance of the semantic space developed here, and another based on the GTM, on a wide range of priming data.

Chapter 5

Simulations

This chapter presents two semantic space models. The first is the high-dimensional model developed in the previous chapter. The second is a low-dimensional map model. We test each model on five semantic priming studies. Two studies were chosen to show how the new models relate to the performance of HAL. The rest exhibit a wide range of priming results that have not been captured before. These include semantic priming for a wide variety of semantic relations with and without association, graded priming and mediated priming phenomena. Mediated priming is of particular interest because HAL has consistently failed to model this effect, and Burgess and colleagues have drawn strong theoretical conclusions about contemporary memory models from the failure.

We first describe the methodology underlying the simulations, before presenting a small example of a low-dimensional topographic lexicon. Simulations using high and low-dimensional semantic spaces follow.

Methodology

There are two distinct ways to interpret semantic space models. A space may be a description of the lexical semantic structure of a language. In this sense, the semantic space described in the previous chapter is a methodology for finding semantic structure in English using a substitutability measure. Alternatively a semantic space may be a theory of semantic representation in people. On the first interpretation when distances in a space correlate reliably with human performance on some psychologically interesting measure we can infer that there is sufficient statistical regularity in the linguistic environment to be able to perform the psychological task. For example

this chapter shows that statistical measures of substitutability in context are sufficient to recreate a wide variety of semantic, associative and mediated priming effects. In the spirit of ecological psychology (Gibson, 1966) and more recently behaviour-based robotics (Brooks, 1991) we might conclude that since we have found the appropriate regularities we simply need to become attuned to them to behave as we do. However, for a computational approach to psychology this is only half the story; there needs to be another theory of how that information is represented in the mind/brain. Semantic spaces *can* be psychological models: e.g. we might assert that each person has vectors of lexical associations and performs similarity computations on them to determine semantic similarity. However, this interpretation is not the one being tested when semantic distances are correlated with a human experimental performance. This is clear from the fact that it when HAL or LSA is compared to human data there is no analysis by subjects, only by items. The work of Finch (1993); Finch and Chater (1994) and Huckle (1996) must be construed in the same way. In the original human experiments there are two sources of variation in the results: variation due to the random sample of subjects chosen for the experiment, and variation due to the random sample of words from the language that are the stimulus materials. In contrast, the simulations studies can only test theories of items, that is, theories of how words are distributionally related in language.

The last chapter developed a new high-dimensional semantic space model; that model should be interpreted as a description of the substitution regularities that underly semantic similarity in English. This chapter develops a new model based on topographic mappings. The topographic map model should be understood as a model of the *representation* of environmental substitution regularities in the brain. This interpretation makes it possible to treat individual maps as subjects in simulation studies and allows analysis by subjects as well as by items.

The chapter considers five experimental studies and attempts to replicate each set of results using a high-dimensional semantic space, and then using a set of topographic map models. If cosines in the high-dimensional model are a good match to the reaction times then we have demonstrated that the substitution relations are in fact sufficient to explain human performance on the study task. If cosines in the low-dimensional map model are a good match to the reaction times then we have also demonstrated that a map representation of semantic space vectors is a plausible model of lexical semantic

representation.

It is important to note that even if the high-dimensional space provides a good match to human data it is not inevitable that the map will also do so. In particular, if the intrinsic dimensionality of the data in semantic space is not in fact very low, the map will fail to extract appropriate structure and the cosines will not match the reaction times. A priori this would seem to be very likely; Landauer and colleagues have suggested that vectors with several hundred elements each are optimal (Landauer and Dumais, 1997). The map representation is equivalent to asking for representative vectors that contain just two elements. Consequently the claim that topographic maps are a good representational model is strong and highly falsifiable.

The introduction of an explicit 'subjects' analysis into psycholinguistic modelling is an important one because it allows the researcher to specify the appropriate cognitive backdrop to the computational processes that are supposed to explain performance. In principal it can move a model closer to being a model of a linguistic agent rather than just a model of an abstract linguistic competence.

5.1 Topographic map models of the lexicon

In these experiments subjects represent noisy versions of co-occurrence information in topographic map form. Topographic maps are ubiquitous in sensory processing (Kandel et al., 1991), and have been considered as the representational substrate for a wide variety of processing tasks (see e.g. Kohonen, 1993, 1995, for a wide range of applications). Ritter and Kohonen (1989) provides an early example of using topographic map models on co-occurrence data. They used a corpus of three word sentences generated from a small number of templates. The application was too small to generate testable psychological predictions but showed how topographic maps might be used to define distributed lexical representation. More recently Miikkulainen (1993, 1997) developed a full natural language understanding system using hierarchies of topographic maps. The input data for the maps was not directly co-occurrence data but rather weights from an augmented backpropagation network (FGREP; see Miikkulainen and Dyer, 1991, for details). Huckle (1996) used an unsupervised neural network to cluster co-occurrence data into semantic classes and compared model predictions with human data. However, the unsupervised network corresponds to the unconstrained mixture model described in Section 2.4 as $\sigma^2 \to 0$. The principal difference between this work

and the application of the GTM described here is in the nature of the assumptions made about co-occurrence data. A mixture -model is guaranteed to be able to fit any data to arbitrary precision if the number of means is allowed to increase without limit (Everitt, 1984). In contrast, this is true of the GTM only if the flexibility of the map is allowed to increase arbitrarily. This corresponds to 'overfitting' due to the loss of the topographic properties of the mapping. For mappings from the latent space into the data space of fixed flexibility, the GTM will only provide reasonable fits to data that is genuinely low-dimensional, irrespective of the number of means at its disposal. Thus a topographic map makes stronger assumptions about the data structure than a standard clustering model. Finch (1993) also uses topographic maps to visualise co-occurrence data, but distances in latent space are not compared with psychological variables.

Topographic maps have also found applications in information retrieval (e.g. Scholtes, 1993; Lin et al., 1991; Kaski et al., 1998; Kohonen et al., 1999). Mathematically these map applications are the most closely related to those developed in this thesis; maps are trained on LSA-style vectors representing each document in a collection. For example, WEBSOM (Kohonen et al., 1999) uses the latent space of a Self-Organizing Map to provide a visualisation of article threads from the newsgroup `comp.ai.neural-nets`. However, despite Anderson's (1983) recommendation that memory access be considered explicitly as an information retrieval problem, these models are not, nor are intended to be, psychological models.

The next section presents a simple example of a lexical representation in a topographic map model (see Lowe, 1997b,a, for further details). The rest of the chapter addresses human experimental results.

5.1.1 An Example

The 'corpus' for this study consisted of sentences sampled from a stochastic context-free grammar described in (Elman, 1990). The grammar was simple enough for the results to be easily interpretable, but contained selectional restrictions appropriate to the semantic classes in the 29 word vocabulary. Each verb subcategorized for semantically appropriate arguments. To mimic the circumstances of data acquisition for larger corpora, punctuation was removed and the entire corpus was concatenated. We used a 10,000 word sample to ensure reliable statistics. It is interesting to note that word frequencies still approximated Zipf's law, even in this highly artificial example. A sample

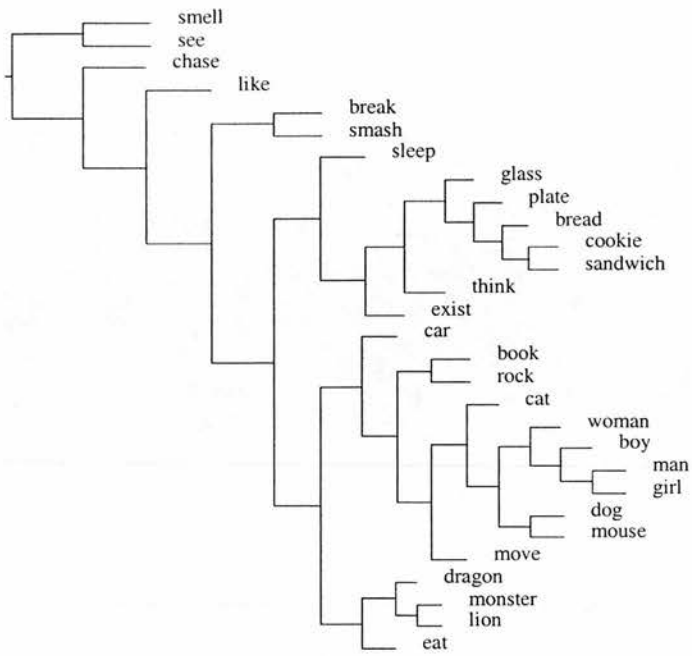

Figure 5.1: A dendrogram for the Elman vocabulary, produced by agglomerative cluster analysis using cosine as a similarity measure.

section of input is shown below.

```
man like boy lion eat mouse
```

Given the small size of the vocabulary (29 words) it was feasible to use all the vocabulary as words and as basis elements. Co-Occurrence statistics were calculated over a window one word either side each word and transformed to log odds-ratios. Figure 5.1 shows the effect of selectional restrictions in the resulting 29-dimensional semantic space.

Figure 5.1 shows that semantically related words cluster in the 29-dimensional semantic space. The map in Figure 5.2 constitutes a low-dimensional model of the data clustered in Figure 5.1 generated by computing the posterior mean for each word in a GTM latent space. The centre of the map corresponds to the latent space vector $[0, 0]$.

The fact that the map preserves semantic distinctions suggests that although the actual dimensionality of the co-occurrence vectors is 29, their intrinsic dimensionality

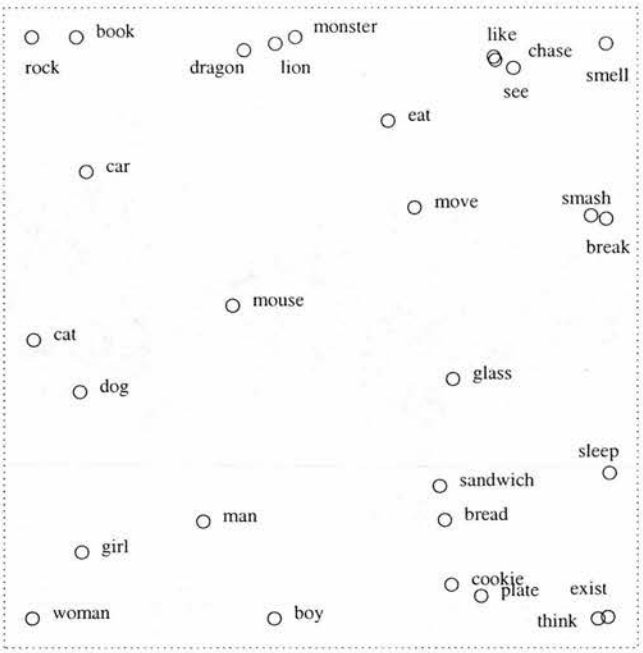

Figure 5.2: Posterior mean positions of each word of the Elman vocabulary in the latent space of the GTM model

is much lower. In the latent space each word clusters with other words that are used in similar contexts; psychological verbs 'see' and 'smell' and 'like' are represented together, as are destructive verbs 'smash' and 'break'. Categorical similarity among the nouns is equally well preserved; human and animal nouns group separately, as do inanimate nouns 'book', 'rock' and 'car'.

As a priming model the posterior means in Figure 5.2 suggest that 'dragon', 'monster' and 'lion' should all prime each other, as should the animate nouns, because priming effects should be proportional to semantic similarity.

Posterior means provide a rather sparse summary of the information present in a map model. With such a drastic dimensionality reduction it is inevitable that some unrelated words are approximately the same distance from a word in the latent space as its 'natural' semantic neighbours. This confound can be resolved by looking at the magnification factor across the map. Items separated by a region of high magnification factor or 'stretch' have been brought together from very different areas of the original data space, whereas those separated by regions of low magnification are genuinely

neighbours in high and low dimensions (see Lowe, 1997b, for details and visualisation). Variable amounts of magnification across the map surface are inevitable when the data are reduced from high dimensions and are almost always present in biological maps; the sensory and motor homunculi in the primary somatosensory and motor cortices are a well-known examples. Magnification is also straightforward to compute in the GTM using the Jacobian of the transformation from latent to data space (Bishop and Williams, 1997) because the mapping is smooth and continuous.

However, at present the use of magnification factors for maps is an heuristic aid to visualisation; it is not yet clear how to factor the effects of magnification into the cosine similarity calculation. Consequently, in the simulations below only the cosine itself is used in order to provide as clean and parameter-independent test as possible, even at the cost of some information loss.

The Elman vectors were used to set an appropriate level of flexibility for the mapping from latent to data space. The map parameters that provided the best visualisation were 1600 evenly spaced latent sample points on $[-1, 1]^2$ and 16 evenly spaced Gaussian basis functions with means one standard deviation apart. To get an idea of the flexibility of the map that was chosen, note that it is intermediate in flexibility between the maps shown in Figures 2.5 and 2.6. These parameters are used for all subsequent simulations.

5.2 Association and Semantic Relatedness

Chapter 3 introduced the conditional probability theory of associative priming, and reviewed the range of semantic relations that Moss and colleagues have shown support semantic priming. Before Moss's work appeared, Shelton and Martin (1992) argued that true semantic priming does not occur without association also being present. They attempted to distinguish the effects of association from those of semantic relatedness in a experiment that compared semantically related pairs with those that were both semantically and associatively related. Facilitation occurred only for the mixed condition.

Lund *et al.* (1995) attempted to replicate Shelton and Martin's findings using HAL with partial success; HAL generated an associative priming effect, but unlike the human experimental results there was also semantic priming of smaller magnitude. Burgess and colleagues argued that the Shelton and Martin's purely associative materials were in fact semantically related, because HAL predicted a priming effect. In itself this is not a very

strong argument since it supports equally the conclusion that HAL actually represents associative relations; Shelton and Martin's materials were carefully chosen to be related, and HAL's distances were not chosen in this sense at all, and so require interpretation. The argument is supported slightly by Burgess and colleagues' observation that only a few of the associative pairs had very short distances between them in HAL, and therefore carried the model's associative effect. Lund *et al.* also argued that elements of the experimental paradigm reduced the priming effect to negligible levels.

HAL was then applied to a larger set of stimuli due to Chiarello et al. (1990). The materials divided into semantically related, associatively related and both semantically and associatively related materials. Human subjects generated robust priming with and without association on these materials, and HAL replicated this performance.

Moss *et al.* (1995; Experiment 2) also showed reliable semantic priming with and without association, in an auditory lexical decision task. However, in their third experiment they also successfully replicated the Shelton and Martin's negative results on the same stimuli using a single word visual lexical decision task. This suggests that experimental paradigm, and perhaps also modality, rather than stimulus characteristics decide the presence of semantic priming without association. However, it is still of considerable theoretical interest to see to what extent semantic space models can generate the priming effects observed in the most expressive paradigm. If semantic spaces generate all the priming effects that have been reported, then they can be put forward as general memory models, and priming failures such as Shelton and Martin reported can be put down to quirks of particular experimental methods.

In experiments 1 and 2 below we apply the high and low-dimensional semantic spaces developed here to Shelton and Martin's materials, both to see whether either space produces results more comparable to the human results, and to compare their performance to HAL. In Experiments 3 and 4 we address the Chiarello materials and compare the results to human performance and to HAL. Finally we model Moss *et al.*'s data, to complete our investigation of the relation between semantic relations, associative relations, and semantic space.

		Related	Unrelated	Effect	Proportion
HAL	Semantic	366	429	63	1.0
	Associated	310	407	97	1.539
Space	Semantic	0.4482	0.2312	0.217	1.0
	Associated	0.5992	0.2673	0.3319	1.529

Table 5.1: Comparison of HAL distances with cosines in semantic space. Effect denotes the absolute difference between unrelated and related means. For HAL this is the average of (unrelated distance - related distance). For the space it is the average of (related cosine - unrelated cosine). Proportion measures how much larger the effect size in the associated condition is compared to the semantic condition.

5.2.1 Experiment 1 : High-dimensional Space Model

Materials and Method

Stimulus materials were taken from Shelton and Martin's (1992) paper investigating semantic priming. In that experiment prime and target word pairs were divided into those that were associatively related and those that were semantically related.

For the purposes of modelling priming, the cosine between a prime and target should be inversely proportional to the corresponding reaction time. The size of a priming effect is calculated by subtracting the cosine between the unrelated prime and target from the cosine between the related prime and target. Cosines for the unrelated prime-target pairs was taken to be the cosine of the target with another prime in the same condition. Cosines are entered directly into analyses of variance.

Results

Table 5.1 shows the distances for HAL and cosines for the semantic space.

There was a reliable effect of overall relatedness, collapsing over the two conditions, $F(1, 70) = 143.104$, $p < .001$, and a main effect of condition, $F(1, 70) = 16.341$, $p < .001$. There was also an interaction, $F(1, 70) = 6.26$, $p < .05$. Relatedness effects were larger in the associated condition.

Priming was reliable in the associated condition $F(1, 35) = 88.947$, $p < .001$, and also in the semantic condition $F(1, 35) = 54.319$, $p < .001$. Table 5.1 shows that the associative priming effect was 1.5 times larger than the semantic priming effect.

Discussion

Table 5.1 shows that HAL and the semantic space are in almost perfect agreement about the relative magnitude of priming effects. However, this is not in complete agreement with the human results which did not show semantic priming.

5.2.2 Experiment 2 : Low-dimensional Model

Method

20 GTM models with the same parameter settings used on the Elman data were trained on 1689 word vectors. The words included all the experimental stimuli presented in this chapter in addition to 1000 filler words of frequency ranks 1000 to 2000 in the BNC (114.55 to 49.15 occurrences per million). Stimulus frequencies ranged from 0.02 to 1639.23 per million, with a median frequency of 33.95 per million.

The addition of large number of irrelevant words was intended to avoid overestimating the intrinsic dimensionality of the data, and also to provide the data matrix with full column rank. The entire augmented set of 534-dimensional vectors was then projected by 20 independently generated stochastic matrices into 50 dimensions (Kaski, 1989). Each randomly mapped data set was used as input for a GTM model.

Neural networks are often criticised for relying crucially on good input representation. Consequently although principal component analysis preprocessing of the vectors would also reduce dimensionality to tractable levels, it would also represent a substantial modelling assumption that is not obviously motivated or interpretable from a neural perspective. Random mapping reduces the dimensionality of the data to a level that is tractable for reasonable network training times while making the fewest possible assumptions about the nature of preprocessing, save that it derives from vectors of lexical associations. Random mapping also introduces variability into the input data

and ensures that no net trains on the same data set. The psychological interpretation of this process is that networks are subjects that have been exposed to roughly the same language data but with significant amounts of noise. We then test the claim that representing this information using topographic maps generates accurate predictions about priming.

Ideally each network would be trained on vectors generated by sampling with replacement from a much larger corpus. However, this is computationally extremely demanding, even were such a corpus available. Using newsgroups is a possible next step in this research. However the BNC and other annotated corpora have a significant advantage over raw text for future work. The work reported here, like that done with HAL or LSA, makes no use of syntactic information. For example, the experimental stimuli 'cup' and 'doctor' have both noun and verb interpretations, but the model treats them as the same word when creating their vectors. But part of speech information can be straightforwardly extracted from corpora using methods very similar to those of semantic space construction. Therefore one logical extension of this work is to make systematic use of this information. Having an annotated corpus allows the semantic space constructor to make use of the detailed syntactic mark-up that comes with corpora. This would not be so straightforward for Usenet text. Also, the BNC is designed to provide a balance of different topics and genres. Usenet discussion is designed *not* to be balanced since it is subdivided into a vast number of groups devoted to specific topics.

To make predictions about priming effects, we project the related prime, unrelated prime and target vectors onto the low-dimensional map surface using Bayes theorem as described in Chapter 2. The map position at the mean of the posterior distribution for each word is a reduced dimension vector representation of the primes and target. We then take cosine measures in the reduced space, just as in the high dimensional model.

Results

Table 5.2 shows the priming results for the low-dimensional model and HAL. There was a reliable effect of overall relatedness, collapsing over the two conditions, $F_1(1, 19) = 920.175$, $p < .001$, $F_2(1, 70) = 39.51$, $p < .001$. The main effect of condition was significant by subjects $F_1(1, 19) = 17.147$, $p < .01$ but not by items, $F_2(1, 70) = 1.996$, $p = .162$. There was no interaction, $F_1 < 1$, $F_2 < 1$.

		Related	Unrelated	Effect	Proportion
HAL	Semantic	366	429	63	1.0
	Associated	310	407	97	1.539
Networks	Semantic	0.6114	0.2667	0.3447	1.0
	Associated	0.6905	0.3494	0.3411	0.99

Table 5.2: Comparison of HAL distances with cosines from 20 networks. Effect denotes priming effect size. Proportion is a standardised measure of priming effect size representing how much larger the associated priming effect size is than the semantic priming effect.

Priming was reliable in the associated condition $F_1(1, 19) = 177.307$, $p < .001$, $F_2(1, 35) = 22.847$, $p < .001$, and also in the semantic condition, $F_1(1, 19) = 776.11$, $p < .001$, $F_2(1, 35) = 17.358$, $p < .001$. Unlike the high-dimensional model the priming effect sizes are almost identical.

Discussion

The low-dimensional model is still less accurate at fitting the human data than the high-dimensional model and HAL, since it loses the relative effect sizes present in the high-dimensional cosines. This appears to be the result of a rather high associated unrelated baseline. Manipulating the unrelated word pairs to stabilise the baseline might bring the low-dimensional results into line with the high-dimensional results, but it is unclear this would help. There would still be a robust semantic priming effect, inconsistent with the human results.

In the next section we consider another experiment designed to be more sensitive to the distinction between associative and semantic relations.

5.2.3 Experiment 3 : High-dimensional model

Lund *et al.* used materials from a hemispheric presentation experiment by Chiarello *et al.* (1990) in a lexical decision priming experiment. The materials were divided into semantically related items, associatively related items, and items that were both semantically and associatively related.

In the human experiment there were main effects of relatedness, and of condition, but no interaction. Most importantly, relatedness effects were present in the semantic condition and in the semantic and associated condition, but there was no associative priming. HAL replicated this pattern of results, so it is interesting to see how the semantic space compares in high and low-dimension.

Results

A comparison of human results, HAL distances and cosines in the semantic space is shown in Table 5.3. There was a main effect of relatedness $F(1, 141) = 155.462$, $p < .001$ and of condition, $F(2, 171) = 4.759$, $p < .01$. This is consistent both with the human results and with HAL. However, there was also an interaction between condition and relatedness $F(2, 171) = 14.39$, $p < .001$. This is because of the stronger effect of association on relatedness.

In the semantic condition there was a reliable effect of relatedness $F(1, 47) = 47.701$, $p < .001$. There were also relatedness effects in the associated condition $F(1, 47) = 13.115$, $p < .01$ and in the associated and semantically related condition, $F(1, 47) = 125.379$, $p < .001$, although the associative priming effect was approximately half the strength of the others.

Discussion

These results are in the same direction as human performance; the associative priming effect is much smaller than in the other two conditions. However, this is not a perfect match to the human results, or to HAL.

Interestingly, the effects of association and semantic relatedness are nearly additive ($0.1845 + 0.1067 = 0.2912 \approx 0.3183$). Ironically this is the associative boost discovered by Moss *et al.*, although it does not appear in the human results. We consider the low-dimensional model next.

		Related	Unrelated	Effect	Proportion
Human	Semantic	643	673	30	1.0
	Associated	623	634	11	0.266
	Both	603	631	28	0.933
HAL	Semantic	347	413	66	1.0
	Associated	322	339	17	0.257
	Both	331	391	60	0.909
Space	Semantic	0.479	0.2945	0.1845	1.0
	Associated	0.397	0.2903	0.1067	0.578
	Both	0.5753	0.257	0.3183	1.725

Table 5.3: Comparison of human results, HAL distances and cosines in semantic space. Effect is the magnitude of the priming effect. For HAL and the human results this is (unrelated - related) and for the space this is (related - unrelated). Proportion is a standardised measure of priming effect size representing how much larger the priming effects are in the associated and the associated and semantically related conditions than in the semantic only condition.

		Related	Unrelated	Effect	Proportion
Human	Semantic	643	673	30	1.0
	Associated	623	634	11	0.266
	Both	603	631	28	0.933
HAL	Semantic	347	413	66	1.0
	Associated	322	339	17	0.257
	Both	331	391	60	0.909
Space	Semantic	0.6986	0.3987	0.2999	1.0
	Associated	0.4208	0.378	0.0428	0.143
	Both	0.678	0.3412	0.3368	1.123

Table 5.4: Comparison of human results, HAL distances and cosines. Proportion is a standardised measure of priming effect size representing how much larger the absolute difference scores are in the associated and the associated and semantically related conditions are than in the semantic priming condition.

5.2.4 Experiment 4 : Low-dimensional model

Results

Results are shown in Table 5.4. There was a main effect of relatedness $F_1(1, 19) = 221.269$, $p < .001$, $F_2(1, 141) = 42.846$, $p < .001$ and of condition, $F_1(2, 38) = 52.403$, $p < .001$, $F_2(2, 141) = 4.044$, $p < .05$. This is consistent both with the human results and with HAL. There was also an interaction $F_1(2, 19) = 78.319$, $p < .001$, $F_2(2, 141) = 7.144$, $p < .01$. This was due to the slight associative boost for semantically related items that are also associated.

In the semantic condition there was a reliable effect of relatedness $F_1(1, 19) =$

240.339, $p < .001$, $F_2(1, 47) = 24.437$, $p < .001$. Priming also occurred in the associated and semantically related condition, $F_1(1, 19) = 293.406$, $p < .001$, $F_2(1, 47) = 33.111$, $p < .001$, but there was no relatedness effect in the associated condition, $F_1(1, 19) = 3.098$, $p = .094$, $F_2(1, 47) < 1$.

Discussion

The low-dimensional model gives a good fit to the human data. In particular the associative priming present in the high-dimensional model has disappeared, and the amount by which association boosts semantic relatedness is much reduced. The predicted level of facilitation for related items in the associated is slightly lower than the human effect sizes. In this case radically reducing dimensionality improves the fit to human data, lending support to the claim that topographic maps are an effective model of semantic representation.

In the next two experiments we consider Moss and colleagues' materials that cross association with semantic relatedness, and at the same time address a wide range of semantic relations. Moss's materials allow us to address all the priming phenomena discussed in Chapter 3, with the exception of mediated and graded priming, which we consider directly afterwards.

5.2.5 Experiment 5 : High-dimensional Model

Moss and colleagues demonstrated semantic priming occurred for all categories of relation, both with and without association. They also showed an interaction between semantic relatedness and association. Semantically related targets were responded to more quickly if they were also associated. This interaction was called the 'associative boost'.

Following the original experimental design we varied three factors: Association (Associated, Non-associated), Semantic Type (Category coordinate, Functional relation) and Relatedness (Related, Unrelated). Semantic Subtypes were nested under Semantic Type: Category coordinates were divided equally into Natural and Artifact object names; Functionally related stimuli were divided between those expressing Instrument and Script relations.

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	697	791	94	724	760	36
Natural	695	804	109	698	755	57
Artifact	699	777	78	750	766	16
Functional	688	759	71	742	783	41
Script	682	762	80	749	780	31
Instrument	695	755	60	735	785	50

Table 5.5: Reaction times from Moss *et al.*'s Experiment 2.

Materials and Method

The semantic space was the same as before. Target words and their related primes were taken from Appendix 1 of Moss *et al.* (1995).

Results

Cosines in the semantic space are shown in Table 5.6. For comparison we reproduce Moss *et al.*'s mean reaction times in Table 5.5.

There was a main effect of relatedness, $F(1, 108) = 1752.534$, $p < .001$, indicating that collapsing over all conditions, semantically related prime-target pairs were more similar than semantically unrelated prime-target combinations. This replicates the semantic priming effect. We found no main effect of semantic type, $F(1, 108) = 1.013$, $p = .09$, and no interaction of semantic type with relatedness, $F(1, 108) < 1$. There was a main effect of association, $F(1, 108) = 33.258$, $p < .001$, replicating the associative priming effects observed in human subjects. There was no associative boost

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.5527	0.3215	0.2312	0.458	0.1806	0.2774
Natural	0.6103	0.2435	0.3668	0.4507	0.2458	0.2049
Artifact	0.4952	0.3996	0.0956	0.4653	0.1153	0.35
Functional	0.5473	0.2592	0.2881	0.3944	0.2408	0.1536
Script	0.5898	0.2754	0.3144	0.3978	0.1991	0.1987
Instrument	0.5049	0.2430	0.2619	0.391	0.2825	0.1085

Table 5.6: Cosines from the high-dimensional semantic space.

$F(1, 108) = 1.571$, $p = .21$, but the interaction between association, relatedness and semantic type was significant, $F(1, 108) = 6.584$, $p < .05$. This is because only among the functional relations did priming effect sizes increase if the items were also associated (see below).

Following the original paper we considered the associated and non-associated items. Following the human results there was a main effect of semantic relatedness in the associated condition, $F(1, 54) = 96.86$, $p < .001$, and no interactions, and in the non-associated condition semantic priming was also present, $F(1, 54) = 85.234$, $p < .001$. There was an interaction with semantic type, $F(1, 54) = 7.041$, $p < .05$. The priming effect for category coordinates was larger than for functional items, which is the same as the human results. This was in part due to a lower cosine baseline for the functional items (0.1806 vs. 0.2408). The subtypes were then examined separately.

Among the category coordinates semantically related pairs were more similar than unrelated pairs, $F(1, 52) = 122.035$, $p < 0.001$. We found no main effects of subtype, $F(1, 52) < 1$, but there was a reliable main effect of association, $F(1, 52) = 25.324$, $p <$

0.001. There was no association × relatedness interaction, $F(1, 52) = 1.008$, $p = .32$.

The functional materials also showed a reliable effect of semantic relatedness $F(1, 52) = 96.544$, $p < .001$ and of association $F(1, 52) = 11.096$, $p < .01$. In contrast to the category coordinates the functional pairs showed a reliable associative boost, $F(1, 52) = 8.96$, $p < .01$. There were no other interactions, although the association × subtype interaction approached significance, $F(1, 52) = 3.549$, $p < .06$.

Separate analyses of the subtypes gave similar results. For the category coordinates, there were reliable effects of semantic relatedness in the natural items, $F(1, 26) = 73.49$, $p < .001$ and in the artifact subtype, $F(1, 26) = 49.233$, $p < .001$. There were also main effects of association $F(1, 26) = 5.569$, $p < .05$ and $F(1, 26) = 22.767$, $p < .001$. Interestingly the association × semantic relatedness interaction was reliable in both subtypes, $F(1, 26) = 5.889$, $p < .05$ and $F(1, 26) = 16.041$, $p < .001$.

For both script and instrument subtypes of the functional condition there were reliable effects of semantic relatedness, $F(1, 26) = 49.446$, $p < .001$ and $F(1, 26) = 49.699$, $p < .001$ The script subtype showed a main effect of association, $F(1, 26) = 13.227$, $p < .01$, but the instrument condition did not, $F(1, 26) = 1.078$, $p = .309$. However the instrument condition did exhibit an associative boost $F(1, 26) = 8.53$, $p < .01$ whereas for the script items the interaction was not reliable, $F(1, 26) = 2.514$, $p = .125$.

Discussion

The high-dimensional model gave results very similar to those found in the original experiment. The model shows robust semantic and associative priming for all semantic types, with and without association. The presence of associative priming is an important result in the light of Burgess and Lund's (1998) claim that since HAL does not reflect associative relations, semantic spaces in general do not.The model also predicts almost the same sized priming effects for category coordinates and functionally related items. This is also consistent with the human results, and supports Moss *et al*'s argument that functional relations are genuinely semantic in nature.

The subtype analyses in the original experiment did not show as much variation as the semantic space, particularly with respect to the unrelated baseline which varies considerably. This may also explain the absence of an associative boost. We test the low-dimensional model next.

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.6212	0.4629	0.1583	0.4847	0.1306	0.3541
Natural	0.7047	0.3556	0.3491	0.4016	0.2553	0.1463
Artifact	0.5376	0.5702	-0.0326	0.5678	0.0059	0.5619
Functional	0.6901	0.403	0.2871	0.4407	0.2613	0.1794
Script	0.7163	0.4234	0.2929	0.2968	0.2366	0.0602
Instrument	0.6639	0.3826	0.2813	0.5847	0.2861	0.2986

Table 5.7: Mean cosine similarity measures from the networks on Moss *et al.*'s data.

5.2.6 Experiment 6 : Low-dimensional model

Method

The GTM models were the same as before.

Results

Mean similarity measures are shown in Table 5.7. There was a main effect of Relatedness, $F_1(1, 19) = 70.884$, $p < .001$, $F_2(1, 108) = 39.752$, $p < .001$, indicating that collapsing over all conditions, semantically related prime-target pairs were more similar than semantically unrelated prime-target combinations. This replicates simple semantic priming. There was a main effect of association, $F_1(1, 19) = 47.125$, $p < .001$, $F_2(1, 108) = 21.411$, $p < .001$, replicating associative priming in human subjects. There was again no associative boost, $F_1 < 1$, $F_2 < 1$, but the interaction between association, relatedness and semantic type was significant by subjects, $F_1(1, 19) = 16.94$, $p < .01$, and marginally significant by items, $F_2(1, 108) = 3.822$, $p = .053$. A

particularly low unrelated baseline for the non-associated category coordinates generated a discernible boost in that condition only.

We looked at the associated and non-associated items separately. There was a main effect of semantic relatedness in both associated and non-associated conditions, $F_1(1, 19) = 89.189$, $p < .001$, $F_2(1, 54) = 14.173$, $p < .001$ and $F_1(1, 19) = 30.527$, $p < .001$, $F_2(1, 54) = 19.509$, $p < .001$. Relatedness $\times$ condition interactions were significant among the associated and non-associated conditions only by subjects, $F_1(1, 19) = 6.88$, $p < .05$ and, $F_1(1, 19) = 9.307$, $p < .01$, respectively. This was because the priming effect was slightly larger among the category coordinates for the associated condition and among the functional items for the non-associated condition.

Category coordinate pairs showed reliable semantic priming, $F_1(1, 19) = 34.126$, $p < .001$, $F_2(1, 52) = 35.019$, $p < 0.001$, and associative priming, $F_1(1, 19) = 28.993$, $p < 0.001$, $F_2(1, 52) = 14.244$, $p < .001$. There was also a reliable associative boost $F_1(1, 19) = 10.824$, $p < .01$, $F_2(1, 52) = 10.824$, $p < .01$, The effect is most striking in the absolute cosine values where related but non-associated pairs take almost the same value as the *unrelated* baseline for associated pairs (see Table 5.7). There was also a significant interaction between association, relatedness and subtype, $F_1(1, 19) = 57.71$, $p < .001$, $F_2(1, 52) = 21.197$, $p < .001$. This was caused by a small negative priming effect among the associated artifacts, and a very large effect in the non-associated condition due to an extremely low unrelated baseline.

Functional items also showed reliable semantic priming $F_1(1, 19) = 106.792$, $p < .001$, $F_2(1, 52) = 15.628$, $p < .001$ and associative priming, $F_1(1, 19) = 42.931$, $p < .001$, $F_2(1, 52) = 7.98$, $p < .01$. The subjects analysis suggested an effect of semantic type $F_1(1, 19) = 5.205$, $p < .05$, that was not maintained across items, $F_2 < 1$. The associative boost was also significant by subjects, $F_1(1, 19) = 5.825$, $p < .05$, but not in the items analysis, $F_2 < 1$.

Among the category coordinates, there were reliable effects of semantic relatedness in the natural items, $F_1(1, 19) = 38.542$, $p < .001$, $F_2(1, 26) = 15.204$, $p < .01$ and in the artifact subtype, $F_1(1, 19) = 24.038$, $p < .001$, $F_2(1, 26) = 20.238$, $p < .001$. and associative priming in natural, $F_1(1, 19) = 23.063$, $p < .001$, $F_2(1, 26) = 4.277$, $p < .05$, and artifact subtypes, $F_1(1, 19) = 19.679$, $p < .001$, $F_2(1, 26) = 12.063$, $p < .001$. The association $\times$ semantic relatedness interaction also occurred in both subtypes. The effect was significant only by subjects in the natural condition, $F_1(1, 19) = 7.47$, $p <$

.05, $F_2(1, 26) = 2.548$, $p = .122$, but was reliable among the artifacts, $F_1(1, 19) = 49.873$, $p < .001$, $F_2(1, 26) = 25.523$, $p < .001$, although this was due to the slight negative priming in the associated group.

Among the functional items, script relations showed semantic priming by subjects, $F_1(1, 19) = 35.775$, $p < .001$, but only approached significance in the items analysis, $F_2(1, 26) = 3.097$, $p = .09$. Instrument relations were more reliable, $F_1(1, 19) = 67.216$, $p < .001$, $F_2(1, 26) = 21.787$, $p < .001$. There was also associative priming for script relations, $F_1(1, 19) = 162.087$, $p < .001$, $F_2(1, 26) = 7.243$, $p < .05$, but not for instruments. Lastly, except for the subjects analysis of script relations there was no significant associative boost.

Discussion

The low-dimensional simulation gave results very similar to those found in the high-dimensional model. Figure 5.7 shows that the relatedness effect size for category coordinates was still very close to the effect size for the Functional items. Associative priming is reliable, but there is still no associative boost.

The difficulty in demonstrating a clear boost in these materials may be due to an unstable baseline. For example, unrelated means in the high-dimensional model vary between 0.1153 and 0.3996 in the artifact subtype. Similarly variable unrelated baseline similarity measures occur in the low dimensional model.

The unrelated primes in both studies were primes from the same condition for a different target word. However, if there is categorical structure in semantic space then we might expect there to be more than random levels of similarity between words in each subtype category. In an attempt to control for this possible confound the experiments were repeated using a set of unrelated primes not contained in the stimulus materials.

5.2.7 Experiment 7 : High-dimensional model

Method

Cosines between related primes and targets were calculated as in the previous experiment. 224 unrelated primes were chosen randomly from the set of padding words presented to the networks in Experiment 2. Cosines were computed between the target vector and a randomly chosen word's vector to generate an unrelated baseline for that

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.5527	0.1733	0.3794	0.458	0.1634	0.2946
Natural	0.6103	0.1593	0.4510	0.4507	0.1620	0.2887
Artifact	0.4952	0.1872	0.308	0.4653	0.1648	0.3005
Functional	0.5473	0.1810	0.3663	0.3944	0.1682	0.2262
Script	0.5898	0.1982	0.3916	0.3978	0.2094	0.1884
Instrument	0.5049	0.1637	0.3412	0.391	0.1270	0.264

Table 5.8: Cosines from the high-dimensional semantic space with unrelated primes chosen randomly from an alternative source.

target. Visual inspection did not reveal any systematic semantic relatedness between randomly chosen words and their targets.

Results

Cosines in the semantic space are shown in Table 5.8. There was a main effect of relatedness, $\mathrm{F}(1, 108) = 314.922$, $p < .001$. We found no main effect of semantic type, $F < 1$, and no interaction of semantic type with relatedness, $\mathrm{F}(1, 108) = 1.303$, $p = .256$. There was a main effect of association, $\mathrm{F}(1, 108) = 16.433$, $p < .001$, replicating associative priming. There was also an interaction between association and relatedness, $\mathrm{F}(1, 108) = 9.939$, $p < .01$. This replicates the associative boost. There was no three way interaction, $F < 1$, and no other significant effects.

We then considered the associated and non-associated items. There was a main effect of semantic relatedness in the associated condition, $\mathrm{F}(1, 54) = 205.972$, $p < .001$, and no interactions. In the non-associated condition semantic priming was also present,

$F(1, 54) = 113.309$, $p < .001$. The priming effect for category coordinates appeared slightly larger than for functional items which is consistent with the human results, but this was not significant, $F < 1$. The subtypes were then examined separately.

Among the category coordinates semantically related pairs were more similar than unrelated pairs, $F(1, 52) = 165.567$, $p < 0.001$. There was also associative priming, $F(1, 52) = 5.607$, $p < 0.05$. There was no associative boost, $F(1, 52) = 2.623$, $p = .111$, and no other significant interactions. The associative boost did not occur due to a low level of similarity between the associated artifact targets and their related primes.

Functional pairs also showed a semantic priming effect, $F(1, 52) = 154.771$, $p < .001$, a main effect of association, $F(1, 52) = 11.555$, $p < .01$, and a reliable associative boost, $F(1, 52) = 8.661$, $p < .01$. There was also a main effect of subtype, $F(1, 52) = 4.58$, $p < .05$. This was due to steadily decreasing amounts of similarity across subtypes relative to a stable baseline (associated related script > associated related instrument > non-associated related script > non-associated related instrument).

Separate analyses of the subtypes gave similar results. For the category coordinates, there were reliable effects of semantic relatedness in the natural items, $F(1, 26) = 83.609$, $p < .001$, and in the artifact subtype, $F(1, 26) = 83.512$, $p < .001$. There was also a main effect of association in the natural subtype, $F(1, 26) = 6.352$, $p < .05$, but this was not present among the artifacts, $F(1, 26) < 1$. The association $\times$ semantic relatedness interaction just missed significance in the natural materials, $F(1, 26) = 4.018$, $p = .056$, but was not present in the artifacts, $F < 1$.

In the functional condition there were main effects of semantic relatedness in the instruments, $F(1, 26) = 114.367$, $p < .001$, and in the script materials, $F(1, 26) = 57.279$, $p < .001$. Both subtypes showed a main effect of association, $F(1, 26) = 5.877$, $p < .05$, and $F(1, 26) = 5.789$, $p < .05$, respectively. Script relations showed an associative boost, $F(1, 26) = 7.03$, $p < .05$, but the instruments did not, $F(1, 26) = 1.859$, $p = .184$. This was due to a slightly lower non-associated than associated unrelated baseline for the instruments.

Discussion

Table 5.8 shows that the unrelated baseline is much less variable than before. Semantic priming is still robust across conditions with and without association. Associative priming occurs reliably among the category coordinates and functional items. However,

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.6885	-0.1511	0.8396	0.4847	-0.2501	0.7348
Natural	0.7047	-0.1634	0.8681	0.4016	-0.1732	0.5748
Artifact	0.5376	-0.1387	0.6763	0.5678	-0.3270	0.8948
Functional	0.6901	-0.1129	0.8030	0.4407	-0.1335	0.5742
Script	0.7163	-0.1023	0.8186	0.2968	-0.0729	0.3697
Instrument	0.6639	-0.1236	0.7875	0.5847	-0.194	0.7787

Table 5.9: Mean cosine similarity measures from the networks on Moss *et al.*'s data with independently chosen unrelated baseline.

there is now a reliable associative boost. This makes the simulation a good model of the human results.

The associative boost is carried by the functional items, since the category coordinates do not produce a significant interaction either together or examined separately by subtype. The delicacy of the boost is demonstrated in the subtype analyses: in the original experiment instruments showed a boost but no main effect of association, whereas in this experiment they show associative priming but no boost. A similar variability of effects was found in the human results.

5.2.8 Experiment 8 : Low-dimensional model

Method

The networks were trained on the same data as before. Unrelated prime vectors were taken from the padding words, rather than from the original stimulus materials, gen-

erating a new unrelated baseline.

Results

Mean similarity measures are shown in Table 5.9. There was a main effect of relatedness, $F_1(1, 19) = 2391.276$, $p < .001$, $F_2(1, 108) = 195.478$, $p < .001$. There was also a reliable effect of association, $F_1(1, 19) = 94.703$, $p < .001$, $F_2(1, 108) = 7.703$, $p < .01$, replicating the associative priming effect. The associative boost was significant by subjects, $F_1(1, 19) = 21.316$, $p < .001$, $F_2(1, 108) = 1.949$, $p = .166$.

There were main effects of semantic relatedness in both associated and non-associated conditions, $F_1(1, 19) = 2358.761$, $p < .001$, $F_2(1, 54) = 103.68$, $p < .001$ and $F_1(1, 19) = 643.53$, $p < .001$, $F_2(1, 54) = 92.124$, $p < .001$.

The category coordinates showed a semantic priming effect, $F_1(1, 19) = 1754.725$, $p < .001$, $F_2(1, 52) = 104.371$, $p < 0.001$, and an associative priming effect, $F_1(1, 19) = 44.774$, $p < 0.001$, $F_2(1, 52) = 4.647$, $p < .05$. No associative boost appeared in either analysis due to the surprisingly large priming effect for non-associated artifacts.

Semantic priming was significant in the functional relations, $F_1(1, 19) = 892.731$, $p < .001$, $F_2(1, 52) = 96.693$, $p < .001$. Associative priming was significant across subjects, $F_1(1, 19) = 47.997$, $p < .001$, and marginally significant in the items analysis, $F_2(1, 52) = 3.403$, $p = .07$, The associative boost was significant for subjects, $F_1(1, 19) = 51.055$, $p < .001$, and approached significance in the items analysis, $F_2(1, 52) = 3.168$, $p = 0.081$.

Separate analyses of the subtypes gave similar results. For the category coordinates, there were reliable effects of semantic relatedness in the natural items, $F_1(1, 19) = 607.939$, $p < .001$, $F_2(1, 26) = 42.241$, $p < .001$ and in the artifact subtype, $F_1(1, 19) = 837.627$, $p < .001$, $F_2(1, 26) = 64.585$, $p < .001$. There were also effects of association for natural items, but only in the subjects analysis, $F_1(1, 19) = 11.81$, $p < .001$, $F_2(1, 26) = 2.592$, $p = .119$. The same pattern held in the artifacts, $F_1(1, 19) = 18.321$, $p < .001$, $F_2(1, 26) = 2.061$, $p = .163$. The association $\times$ semantic relatedness interaction also occurred only in subjects analyses for both subtypes, although the trend is visible.

Among the functional items, script relations produced reliable semantic priming $F_1(1, 19) = 258.341$, $p < .001$, $F_2(1, 26) = 35.26$, $p < .001$, as did the instruments $F_1(1, 19) = 958.256$, $p < .001$, $F_2(1, 26) = 63.229$, $p < .001$. Associative priming for

script relations was significant by subjects, $F_1(1, 19) = 78.521$, $p < .001$, and marginally significant by items, $F_2(1, 26) = 3.205$, $p = .08$. Associative priming for instruments was significant only by subjects, $F_1(1, 19) = 4.98$, $p < .05$, $F_2 < 1$. Script relations showed a reliable associative boost, $F_1(1, 26) = 45.579$, $p < .001$, $F_2(1, 26) = 6.438$, $p < .05$, but did not.

Discussion

The low-dimensional simulation gave results rather similar to those found in the original experiment, particularly in the subjects analysis. The Relatedness effect size for Category Coordinates was still very close to the effect size for the Functional items. Semantic and associative priming still occurred although the associative boost was not reliable.

5.2.9 General Discussion

In experiments 1 and 2 we have seen that the high and low-dimensional models agree with HAL in predicting more associative priming than observed in human behaviour on Shelton and Martin's materials. We have also seen in Experiments 3 and 4 that a more carefully factorized set of materials due to Chiarello and colleagues allows the spaces to accurately predict human behaviour. Experiments 5 to 8 show that another set of carefully factorized materials generate priming effects that can be captured by the semantic spaces presented here. It is clear then that the spaces can be excellent predictors of associative and semantic priming, and also the interaction between them demonstrated by Moss *et al.* However, the nature of association is still extremely unclear.

Shelton and Martin's materials generated a large associative priming effect and negligible semantic facilitation. In contrast, Chiarello *et al.*'s materials generated a large semantic only effect and negligible associative priming. And Moss *et al.*'s materials demonstrated a pattern in between these two extremes. These results hint that, although the free association task sometimes generates word pairs that prime, it often does not; that is, the underlying *cause* of relatedness does not match the free association task tightly.

The conditional probability theory of associative priming (to which Moss and co-authors subscribe) postulates a processing mechanism that is sensitive to conditional

probabilities of word occurrence in text. If 'pan' is more likely than other words to appear soon after 'bed' in text then, according to this theory the two words will become associated. In support of the conditional probability theory, Spence and Owens, 1990 showed that in the Brown corpus, associatively related words tended to occur within 250 characters (approximately 50 words) of each other significantly more often than non-associatively related words.

Conditional probability is a plausible candidate for the underlying reason why words are related. Neural network instantiations of the conditional probability theory (Plaut, 1995; Moss et al., 1994) force the network to learn to map orthographic or phonological representations of words onto semantic features on the basis of training sequences that have been manipulated to exhibit the correct occurrence probabilities. There are then two types of information – semantic relatedness between words which is represented explicitly in semantic features, and associative relatedness which depends on conditional probability and is represented implicitly in the network parameters. In constrained computational domains the conditional probability theory of association generates the right predictions. However, this does not explain why words that are frequently generated in a free association task sometimes do and sometimes do not generate reliable priming effects. Surely if the conditional probability of 'pan' is high given 'bed' then 'pan' will be likely to be generated by 'bed' in free association.

The substitutability theory of association

In contrast, neither the high nor the low-dimensional semantic space models make the distinction between semantic features and conditional probability. The success of the spaces suggests that it is not in fact necessary to postulate a distinct form of information or an additional processing mechanism to explain associative priming; semantically and associatively related words are simply more substitutable in context than unrelated words. This alternative theory, call it the substitutability theory of association, provides a more parsimonious account of associated priming effects.

The theory explains why we find many experimental studies where associative pairs are clearly semantically related according to a semantic space. Ultimately, the problem of understanding the nature of semantic and associative relations might be better served by reformulating in terms of substitutability measures and conditional probability.

It is important to see that the conditional probability theory need not necessarily

contradict a substitutability account. This can be seen more easily by examining the process of computing word vectors. If 'cup' significantly raises the probability of seeing 'saucer' then we may expect them to occur near each other in text. This is the fact that drives neural network models of associative priming. However it also means that the windows over which the vectors for 'cup' and 'saucer' are calculated will substantially overlap. The higher the conditional probability of 'saucer' given 'cup' then the more often we expect to see them together and the smaller the gap we expect to see between them. This means that vector elements for word pairs related by high conditional probability will be computed from 'shared' counts with the result that they will be more numerically similar than those generated from independent occurrences. For example, with a window size of 5 words, the textual fragment

```
he put the cup by the plate because its matching saucer had been
broken
```

adds 1 to f(plate, cup) and one to f(plate, saucer) at the same time. Because estimates of lexical association depend on co-occurrence counts, the more often this count-tying occurs the closer the estimated association between 'cup' and 'plate' becomes to the estimated association between 'saucer' and 'plate'.

We have shown that the semantic space represents associated words as more substitutable in context above. We can also confirm that the analysis above actually holds for the Moss stimuli. If the probability of a target occurring occurring within the window of an associatively related prime is higher than for a non-associatively related prime then the analysis is vindicated. McDonald and Lowe (1998, Experiment 2) compared the probability of each target given its associated prime, to its probability given a non-associated prime and found that associatively related primes were indeed significantly more likely to occur within 3 words of their targets than non-associated primes according a Mann-Whitney test, $U = 630$, $p < .001$. The window size for this experiment was only 3 words either side, so this is a particularly stringent test of the analysis (see paper for details of the experiment).

One interesting aspect of this analysis is that we know that many associatively related words e.g. 'bed' and 'pan' *cannot* be substituted for one another in context. However, the standard methods for estimating substitutability employed by a semantic space does not distinguish between words that are tight collocates but not substitutable, and words that are genuinely substitutable. This 'failure' provides a good model of

associative priming, and also suggests a more parsimonious explanation of associative relatedness.

5.3 Graded and Mediated Priming

The previous experiments have shown that the semantic space model and its low-dimensional version can account for the detailed structure of human semantic and associative priming behaviour. However, apart form the Moss experiment, the high-dimensional version has not shown considerable difference from HAL. In the next two experiments we model graded and mediated priming. This is interesting for two reasons. First, the existence of mediated priming has been considered to be a crucial piece of evidence for the spreading activation theory of semantic memory. Second, the HAL model has been shown not to be able to model mediated priming. Therefore, if it can be modelled with the semantic spaces developed here, the two models will have been sharply distinguished.

Mediated priming has been put forward as a crucial test for theories of semantic memory (Neely, 1991). According to spreading activation theory (e.g. Anderson, 1983), when a word is presented it activates its representation in a network structure in which semantically related words are directly connected; more generally, the semantic similarity of two words depends on the number of links that must be traversed to reach one to the other. The level of activation controls the amount of facilitation received by the corresponding word. Although ultimately every word can be reached from any location in the network, activation decays during memory access so only a few of the most related words are facilitated. Spreading activation theories predict that a prime word should facilitate a target word directly as described above, for example when "tiger" facilitates "stripes" for pronunciation or lexical decision. Spreading activation theory also predicts that "lion" will facilitate "stripes" when activation spreads from the representation of "lion" to that of "stripes", via the related concept of tiger (de Groot, 1983; Neely, 1991).

Small but reliable mediated priming has been demonstrated for pronunciation tasks but is less reliable for lexical decision (Balota and Lorch, 1986). Spreading activation theory explains the size of the priming effect by arguing that "lion" and "stripes" are only indirectly related in semantic memory so that activation has decayed significantly by the time activation from `lion` reaches `stripes`.

Theories that do not assume the existence of activation or a network structure in semantic memory, e.g. compound cue theory (Ratcliff and McKoon, 1988; McKoon and Ratcliff, 1998), cannot take advantage of either of the priming explanations above. In compound cue theory, direct priming is explained roughly as follows: the prime and target are joined in a compound cue that is compared to representations in long-term memory. The comparison process generates a 'familiarity' value which controls the size of the priming effect. The essential feature of this explanation is that, unlike spreading activation theory, there is no mention of the intermediate representation "tiger" when explaining how "lion" facilitates "stripes". But it is less clear how compound cue theory should explain mediated priming.

In response to this difficulty, McKoon and Ratcliff (1992) have argued that the mediated priming effect is not in fact due to activation spreading through an intervening representation, but is due to direct but weak relatedness between the prime and target words. To address the issue of priming effect magnitude they provided a quantitative method for generating prime target pairs with various degrees of relatedness. The method is based on pointwise mutual information (Church and Hanks, 1990) computed over a corpus. McKoon and Ratcliff's (1992) Experiment 3 showed that their method produced stimuli that reliably generate a range of priming effect sizes, and that the sizes can be controlled. They then argue that mediated priming is simply a special case of graded priming.

Livesay and Burgess (1998b,a) replicated the mediated priming effect in human subjects using a pronunciation task, but had less success with lexical decision (the same situation that was reported in Balota and Lorch's original paper). In an attempt to understand the nature of the priming mechanism they found that mediated primes from the Balota and Lorch stimuli could be divided heuristically into contextually appropriate and contextually inappropriate word pairs. Subsequent analysis revealed that only contextually appropriate pairs generated a priming effect. They then compared distances between each type of prime (direct or mediated) and their targets in HAL, a semantic space model (Lund et al., 1995). Burgess and colleagues have argued that distances in HAL reflect semantic relatedness; shorter distances are argued to reflect greater semantic relatedness (see Burgess et al., 1998, e.g.). Directly related primes were on average closer to their targets than the corresponding unrelated primes, so HAL successfully replicated the direct priming effect. However, both contextually appropriate

and contextually inappropriate mediated primes were *further* from their targets than unrelated controls. Thus distances in HAL predict that the mediated primes should slow responses to their targets, relative to an unrelated baseline. Subsequent analysis showed that for contextually consistent primes, greater distance correlated 0.6 with greater priming effects.

Livesay and Burgess concluded that mediated priming could not be due to a direct but weak relatedness effect between mediated primes and their targets on the grounds that HAL predicted the wrong effect. They then explored the possibility, suggested by McKoon and Ratcliff's paper, that mediated priming is determined by raw co-occurrence frequency between prime and target but found no significant correlations.

The following experiments model human performance on the stimuli generated by McKoon and Ratcliff using pointwise mutual information. We refer to these stimuli as the mutual information stimuli. These results demonstrate that McKoon and Ratcliff's direct theory of mediated priming is at least consistent with explanations of priming based on semantic space. The next set of experiments tackles mediated priming directly by replicating the results of Livesay and Burgess's mediated priming experiment.

5.3.1 Experiment 9 : High-dimensional model

Materials

Stimuli for this experiment are the same as those used in McKoon and Ratcliff's Experiment 3. They are word quadruples of the form ⟨free-association prime, high-t prime, low-t prime, target⟩ . Free association primes were chosen from association norms. High and low-t primes were chosen by first calculating a measure of lexical association based on the T-statistic between each target word and a large number of candidate primes (Church and Hanks, 1990). McKoon and Ratcliff divided the candidate primes for each target into those with high values of the T-statistic (high-t primes) and low values (low-t primes). Unrelated primes were related primes from another quadruple.

McKoon and Ratcliff's subjects responded fastest to target words preceded by an associated prime, next fastest to a high-t prime, slower to a low-t prime and slowest of all to an unrelated prime (see Table 5.10, line 1). There were priming effects for associated pairs and high-t pairs, but the low-t group was not significantly different from the unrelated baseline.

	Related	High-t	Low-t	Unrelated
M&R	500	528	532	549
Space	0.5475	0.3644	0.3043	0.2179

Table 5.10: Mean reaction times in msec. from McKoon and Ratcliff and cosines in semantic space for the mutual information stimuli.

Results

Mean cosines and reaction times for the mutual information stimuli are shown in Table 5.10. There was a main effect of condition, $F(3, 117) = 29.942$, $p < .001$. There was also a reliable associative priming effect, $F(1, 39) = 90.193$, $p < .001$, and High-t pairs were significantly more similar than the unrelated baseline, $F(1, 39) = 17.253$, $p < .001$. Low-t pairs were also reliably more related than the unrelated baseline, $F(1, 39) = 6.919$, $p < .05$, although the effect size was considerably smaller than in other conditions.

Discussion

The semantic space successfully replicates the graded nature of McKoon and Ratcliff's stimuli. The only slight difference between the two is that the low-t group in the human study were not significantly faster than the unrelated group, although a trend was the clearly visible.

The experiment also provides some interesting tangential support for the conditional probability theory of association above. The high and low-t groups were chosen using a method that maximises lexical association in text; low-t and high-t primes should be increasingly more likely to occur near the target, taking chance into account. The difference between the conditional probability theory is only that chance has been

	Related	High-t	Low-t	Unrelated
M&R	500	528	532	549
Networks	0.5489	0.3291	0.2966	0.2073

Table 5.11: Mean reaction times in msec. from McKoon and Ratcliff and mean cosine similarity measure for 20 networks trained on the mutual information stimuli.

factored in; The analysis of Chapter 4 suggests that a lexical association function that takes chance into account will make the same predictions as simple conditional probability estimates when the occurrence frequencies of target and prime happen to be the same. This is generally the case for experimental stimuli. Thus the success of the space suggests that steadily increasing conditional probability is sufficient to cause graded similarity in semantic space.

5.3.2 Experiment 10 : Low-dimensional model

Method

The networks were the same as before.

Results

Table 5.11 shows very similar results for the low-dimensional model. There was a main effect of condition, $F_1(3, 57) = 55.811$, $p < .001$, $F_2(3, 117) = 5.437$, $p < .01$. There was also an associative priming effect, $F_1(1, 19) = 171.812$, $p < .001$, $F_2(1, 39) = 18.084$, $p < .001$, and a priming effect for the high-t group, $F_1(1, 19) = 19.2$, $p < .001$, although this was not significant for items, $F_2(1, 39) = 1.599$, $p = .214$. The priming effect for low-t materials was only significant by subjects $F_1(1, 19) = 11.831$, $p < .01$, but not by items $F_2(1, 39) = 1.08$, $p = .3$.

Discussion

The low-dimensional results are very similar to those of the high-dimensional model, although slightly less reliable. The graded nature of priming is still accurately reflected in the cosine measures.

These experiments show that graded semantic priming can be captured in detail in the semantic space, whether in high or low-dimensional formulation. Also, McKoon and Ratcliff's use of an alternative lexical association measure for choosing related primes makes it possible to see what the effects are on the semantic space. Empirically it appears that the alternative measure used to choose these stimuli is perfectly compatible with a semantic space model based on substitutability. This is explained by the substitutability theory of association. We consider mediated priming next.

5.3.3 Experiment 11 : High-dimensional model

In the pronunciation task both Balota and Lorch and Livesay and Burgess's subjects showed direct and mediated priming (see Table 5.12, lines 1 and 2). Mediated priming effects were smaller than direct priming effects.

Materials and Method

Stimuli were word triples of the form, ⟨mediated prime, directly related prime, target⟩ taken from Balota and Lorch's (1986) paper. One triple had to be discarded due to very low frequency in the corpus. A randomly chosen triple was discarded from each of the other two prime conditions to maintain balance. The semantic space is the same as before.

Results

Mean reaction times for Balota and Lorch's subjects and Livesay and Burgess's subjects are shown with cosines in semantic space in Table 5.12. The prime conditions were significantly different $F(2, 88) = 18.844$, $p < .001$ and we performed pairwise analyses of variance to examine the differences in more detail.

There was a reliable direct priming effect (0.212 vs. 0.085), $F(1, 44) = 24.724$, $p < .001$ and also a reliable mediated priming effect of smaller magnitude, $F(1, 44) = 15.635$, $p < .001$.

	Related	Mediated	Unrelated
B&L Pron.	549	558	575
L&B Pron.	576	588	604
Space	0.212	0.164	0.084

Table 5.12: Mean reaction times in msec. for the pronunciation experiments of Balota and Lorch (B&L, line 1) and Livesay and Burgess (L&B, line 2) in msec. Cosine measures for the same materials are on line 3.

Discussion

The high-dimensional model models the mediated priming effect accurately, annlthough the space's mediated effect size is slightly larger than the equivalent human result. Since there is no method of mediation in semantic space, only varying amounts of substitutability, we may conclude that Livesay and Burgess are incorrect to claim that a semantic space cannot generate a mediated priming effect.

These results also show that mediated priming cannot be used to distinguish spreading activation accounts from compound cue models. Substitutability between words is precisely the direct measure that Ratcliff and McKoon suggests underlies mediated priming. However, not only are there no nodes for activation to spread between in a space, but there are also no cues to combine and compare against long term memory stores. The high-dimensional semantic space therefore provides an explanation of mediated priming that is more parsimonious than both traditional models. We consider the low-dimensional models next.

	Related	Mediated	Unrelated
B&L Pron.	549	558	575
L&B Pron.	576	588	604
Maps	0.3968	0.3549	0.3673
RM	0.3876	0.3262	0.2664

Table 5.13: Mean reaction times for the pronunciation experiments of Balota and Lorch (B&L, line 1) and Livesay and Burgess (L&B, line 2) in msec. Similarity measures for networks are on line 3 and for the random mapping only on line 4.

5.3.4 Experiment 12 : Low-dimensional model

Method

The GTM models were the same as before.

Results

Similarity measures and reaction times are shown in Figure 5.13. The prime conditions were not significantly different $F_1(2, 38) = 2.467, p = 0.089, F_2 < 1$, and pairwise analyses of variance revealed no direct priming effect, $F_1 < 1, F_2 < 1$ and no mediated priming effect $F_1(1, 19) = 1.74, p = .2, F_2 < 1$. Indeed the mediated pairs were *less* similar than the unrelated pairs, though this difference was not significant.

Discussion

These results are not comparable to the high-dimensional model described above. There was no direct or mediated priming effect generated by the networks. Interestingly, the

direction of mediated priming seemed to be reversed so that mediated pairs were less similar and predicted longer reaction times than unrelated pairs. Although this effect was not statistically significant, it is strikingly similar to Livesay and Burgess's results on the same mediated priming stimuli.

There are a number of reasons the low-dimensional model could fail to generate mediated priming. Perhaps the weaker relations between mediated priming stimuli are not well captured by a very low-dimensional projection of the high-dimensional space. They may be intrinsically higher-dimensional. This would suggest that at least some semantic priming effects cannot be captured by low-dimensional projection, and that a complete theory of priming in semantic memory requires a high-dimensional space.

Alternatively the random mapping may be distorting the geometric relationships in semantic space too much for the map to pick out relevant structure. To investigate the possible distorting role of the random mapping we ran the same analyses on the randomly mapped vectors as we had done on the posterior means from the networks.

5.3.5 Experiment 13 : Random-mapping only

Materials and Method

Semantic space vectors for 135 mediated priming stimuli taken from Experiment 12 were subjected to random mapping into 50 dimensions. 50-dimensional vectors were used in the same way as posterior mean values.

Results

The results of using only randomly mapped vectors were essentially identical to the high-dimensional model. Mean similarity values for the randomly mapped vectors are given in line 4 of Table 5.13. The prime conditions were significantly different $F_1(2, 38) = 359.07$, $p < .001$, $F_2(2, 88) = 13.074$, $p < .001$. There was also a clear direct priming effect, $F_1(1, 19) = 546.149$, $p < .001$, $F_2(1, 44) = 20.486$, $p < .001$, and a reliable mediated priming effect, $F_1(1, 19) = 173.034$, $p < .001$, $F_2(1, 44) = 5.295$, $p < .05$.

Discussion

The result above suggest that there is sufficient information in random projections of the 536 dimensional space into 50 dimensions to model the mediated priming effect. Both direct and mediated effect appear in this model.

Experiment 13 is in fact only one of a sequence of studies investigating the degree to which vectors for the mediated priming stimuli could be distorted and still generate the correct psychological predictions. The complete study consisted of 10 individual studies in which the target mapping dimensionality was reduced from 50 to 10 in steps of 5. For each target dimensionality we recomputed the corresponding random mapping and submitted cosines between the vectors to the ANOVAs above. Mediated priming effects remained reliable down to dimensionality 15. This supports the claim that the true dimensionality of the high-dimensional data is quite low, although it does not explain why the networks did not pick it out.

We also considered the possibility that there there was not enough information available in 50 dimensions for the networks, even if simple distances in this space did generate the correct psychological prediction. In Experiment 14 we investigate the effects of increasing the target dimensionality of the randomly mapped high-dimensional vectors.

5.3.6 Experiment 14 : Increased input dimensionality

Materials and Method

536-dimensional vectors for the mediated priming stimuli were mapped randomly into a 100 dimensional space and presented to the 20 networks for training.

Results

Mean similarity values are shown in Table 5.14. Differences between prime conditions attained significance in the subject analyses $F_1(2, 38) = 5.887$, $p < .001$, but not for items, $F_2(2, 38) = 0.472$, $p = .62$. Similar patterns of results held for pairwise ANOVAs: Direct priming effects were significant by subjects, $F_1(1, 19) = 8.299$, $p < .01$ but not by items, $F_2 < 1$. Mediated priming effects were not significant in either analysis.

	Related	Mediated	Unrelated
B&L Pron.	549	558	575
L&B Pron.	576	588	604
Networks	0.5717	0.5233	0.5175

Table 5.14: Mean reaction times for the pronunciation experiments of Balota and Lorch (B&L, line 1) and Livesay and Burgess (L&B, line 2) in msec. Similarity measures for networks trained on 100-dimensional projections of original high-dimensional vectors are shown on line 3.

Discussion

In this experiment the data in Table 5.14 show the correct trend (related > mediated > unrelated). However, priming is still not statistically reliable in the items analysis.

5.4 Conclusion

This chapter has presented high and low dimensional semantic space models and tested them on experimental results from a wide range of semantic priming experiments.

Experiments 1 to 8 show that both high and low-dimensional models succeed in generating semantic and associative priming effects of the same or similar magnitude to human subjects. In Experiments 7 and 8 we also capture an interesting interaction between these two type of relatedness.

In contrast to previous recurrent networks models of the relation between association and semantic relatedness, neither the high nor the low-dimensional models have a separate mechanism for detecting conditional probability relations between words. However, their success in predicting priming effects regardless of the absence moti-

vates an alternative to the theory that association between words depends on high conditional probability. The substitutability theory of association explains how simple measures of substitutability in context implemented in a semantic space are affected by changes in conditional probability, and lead to conditionally probable word pairs being judged as more substitutable. The substitutability theory provides a more parsimonious explanation of associative priming because it makes no assumptions about mechanism.

In the remaining experiments we show that graded priming effects can also be captured in a semantic space, despite the fact that the stimuli for these experiments are generated by a process similar to measuring conditional probability between words. The success of the models on these stimuli also lends support to the substitutability theory.

Mediated priming has been argued to be a crucial test between spreading activation and compound cue models of priming. Compound cue theories have no mechanism of mediation, therefore if mediated priming relies on spreading activation in a network structure then these theories should fail to capture the effect. However, compound cue theorists have argued that mediated priming is due to weak but direct relatedness. If this is true then a semantic space should capture the effect by predicting less priming for mediated word pairs than for directly related pairs. HAL does not predict this so Burgess and colleagues conclude that mediated priming is not due to direct relatedness. However, we have seen that an appropriately constructed semantic space generates exactly the prediction compound cue theorists require, and successfully models mediated priming as a special case of graded priming effects. On the other hand the semantic space is a parsimonious explanation of mediated priming because it makes no architectural assumptions at all. The upshot for psycholinguistic models is that mediated priming cannot be used to distinguish between the spreading activation and compound cue architectures because a parsimonious explanation is available that makes neither kind of architectural assumption.

Curiously the low-dimensional model fails to predict mediated priming. However, we showed that very low dimensional projections of the semantic space vectors did generate mediated priming, and that networks that did not suffer as drastic a preprocessing on their input data generate the correct predictions, although this was not statistically reliable. This issue merits further investigation.

We may conclude that the success of these models in predicting a wide range of semantic priming effects suggests that the high-dimensional semantic spaces developed in previous chapters are more successful than other semantic space models. The success of the low dimensional models supports the hypothesis that the intrinsic dimensionality of semantic space is very low, and may be effectively represented by topographic maps.

Chapter 6

Conclusions

The previous chapters have developed two novel semantic space models and compared their performance to human experimental data and to alternative models:

Chapter 2 introduced latent variable models for real valued data, developed their neural interpretation and presented a unifying review of topographic mapping models in the neural and statistical literature. We described how familiar linear Gaussian models such as Factor Analysis can be extended by introducing smooth non-linear mappings to produce statistical models that double as flexible and statistically interpretable topographic maps, and showed how a Gaussian Process formulation of Bishop *et al.*'s Generative Topographic Mapping model constituted a full probabilistic model of topographic mapping to which a wide range of current map models approximate. Following the Generative Turn in computational neuroscience and neural networks we described activity in a topographic map as the posterior probability of position in a latent lower-dimensional data space.

Chapter 3 introduced the basic data of semantic priming and described how spreading activation, compound cue, and neural network models dealt with them. Concentrating on the statistical aspects of neural networks leads to two conclusions about network models of semantic memory. First, Masson's Hopfield network model of semantic memory must be impractical as a memory model because the number of lexical entries it can store degrades rapidly as correlations between semantic features increase. However, any useful semantic representation, whether based on features or functions of co-occurrence counts must introduce substantial correlations between words otherwise it cannot be said to express what relates them. Second, a probabilistic analysis of re-

current networks that perform a one-way mapping from phonology or orthography onto semantic representations shows that they cannot be simply reversed to model naming rather than lexical decision. They are thus fundamentally limited as psycholinguistic models.

Chapter 4 showed how semantic space models relate to statistical generalisations of the replacement test from theoretical linguistics. This understanding motivated a set of new statistically-motivated construction methods for space models. We introduced a new measure of lexical association in text that takes chance co-occurrence into account and compared it to previous work in this area. One particularly important result is that raw co-occurrence counts have a strong frequency bias due to the distributional properties of text, described by Zipf's law. Even distributionally independent, and therefore semantically unrelated words of differing frequencies can be expected to generate different co-occurrence depending on their occurrence frequencies. Therefore any distance measure that uses raw counts will only give correct results for equally frequent words. A version of this problem makes it clear why choosing context words as the HAL model does is not to be recommended. An alternative method for choosing context words is presented that treats context words as raters and judges their reliability over corpus sections. We also showed how LSA related to the statistical latent variable models presented in Chapter 2 using a reformulation as principal component analysis.

Chapter 5 tested the high-dimensional semantic space developed in Chapter 4, and a low-dimensional model based on topographic mapping of noisy semantic space vectors, on a range of empirical data. We addressed the relation between associative and semantic priming in Experiments 1 to 8, showing that both high and low-dimensional models succeed in generating semantic and associative priming effects of the same or similar magnitude to human subjects. We also described an alternative to the conditional probability theory of association, called the substitutability theory, that explained how semantic space models captured associative priming effects without having any method of tracking conditional probabilities. This is a new theory of associative priming that is more parsimonious than the conditional probability version, and yet explains all the data that theory can.

Experiments 7 and 8 showed that high and low-dimensional models also capture a subtle interaction between semantic relatedness and association discovered by Moss and colleagues. In the same experiments we show that both models predict priming for

a wide range of semantic relations, consistent with human data.

Experiments 9 to 14 investigated graded and mediated priming. We showed that graded priming could be accurately modelled, even when the stimuli were only associatively related, or were generated by a method designed to find word pairs with high levels of lexical association, not substitutability. Mediated priming has been argued to be a crucial test between spreading activation and compound cue models of priming. Since compound cue theories have no mechanism of mediation, if mediated priming relies on spreading activation in a network structure then these theories should fail to capture the effect. However, compound cue theorists have argued that mediated priming is due to weak but direct relatedness. If this is true then a semantic space should capture the effect by predicting less priming for mediated word pairs than for directly related pairs. HAL does not predict this so Burgess and colleagues have concluded that mediated priming is not due to direct relatedness, and that semantic spaces cannot model the effect. However, we showed that an appropriately constructed high-dimensional semantic space generates exactly the prediction compound cue theorists require, and successfully models mediated priming as a special case of graded priming effects. On the other hand the semantic space is a parsimonious explanation of mediated priming because it makes no architectural assumptions at all. We concluded that mediated priming cannot be used to distinguish between the spreading activation and compound cue architectures because a parsimonious explanation is available that makes neither kind of architectural assumption. Although the low-dimensional model fails to predict mediated priming, we showed that very low dimensional projections of the semantic space vectors did generate mediated priming, and that networks that did not suffer as drastic a preprocessing on their input data generate the correct predictions, although not yet reliably.

The successful modelling of semantic priming for a wide variety of semantic relations, for associative priming, for graded priming and for mediated priming suggests, first, that substitutability in context, implemented as a semantic space model, is an extremely powerful representational medium for lexical semantic information, and second, that the intrinsic dimensionality of semantic space data is in fact extremely low and can often be accurately approximately in topographic form.

Bibliography

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley and Sons.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley and Sons.

Anderson, J. R. (1976). *Language, Memory and Thought.* Lawrence Erlbaum Associates.

Anderson, J. R. (1983). *The Architecture of Cognition.* Harvard University Press.

Atsugi, A. (1998). Density estimation by mixture models with smoothing priors. *Neural Computation*, 10(8):2115–2135.

Balota, D. A. and Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning Memory and Cognition*, (12):336–345.

Barbosa, P., Fox, D., Hagstrom, P., McGinnis, M., and Pesetsky, D., editors (1998). *Is the best good enough: Optimality and competition in syntax*, MIT Working Papers in Linguistics. MIT Press.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Batali, J. (1998). Computational simulations of the emergence of grammar. In Hurford, J. R., Studdert-Kennedy, M., and Knight, C., editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 405–426. Cambridge University Press.

Batali, J. (2000). Negotiating syntax. to appear in Proceedings of the 3rd Conference on the Evolution of Language.

Bell, T. and Sejnowski, T. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, (7):1004–1034.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235.

Bishop, M. S. C. M. and Williams, C. K. I. (1997). Magnification factors for the SOM and GTM algorithms. In *Proceedings of the 1997 Workshop on Self-Organizing Maps, Helsinki, Finland.*

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Addison–Wesley.

Brooks, R. A. (1991). Intelligence without reason. Memo 1293, MIT Artificial Intelligence Laboratory.

Bullinaria, J. A. (1995). Modelling lexical decision: Who needs a lexicon? In Keating, J. G., editor, *Neural Computing Research and Applications III*, pages 62–69.

Bullinaria, J. A. and Huckle, C. C. (1997). Modelling lexical decision using corpus derived semantic representations in a connectionist network. In Bullinaria, J. A., Glasspool, D. W., and Houghton, G., editors, *4th Neural Computation and Psychology Workshop*, pages 213–226. Springer Verlag.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25):211–257.

Burgess, C. and Lund, K. (1996). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, (12):177–210.

Burgess, C. and Lund, K. (1998). The dynamics of meaning in memory. In Dietrich, E. and Markman, A., editors, *Cognitive Dynamics: Conceptual Change in Humans and Machines.* under review.

Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora.*

Caid, W. R., Dumais, S. T., and Gallant, S. I. (1995). Learned vector space models for information retrieval. *Information Processing and Management*, 31(3):419–429.

Cann, R. (1996). Categories, labels and types: Functional versus lexical. Edinburgh Occasional Papers in Linguistics EOPL-96–3, University of Edinburgh.

Chiarello, C., Burgess, C., Richards, L., and Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain and Language*, (38):75–104.

Christiansen, M. H. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23:157–205.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, (16):22–29.

Collins, A. M. and Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, (82):407–428.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, (8):240–248.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge University Press.

de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal Learning and Verbal Behavior*, pages 417–436.

Deese, J. (1965). *The Structure of Associations in Language and Thought.* Johns Hopkins Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dunning, T. (1993). Accurate methods for the statistics for surprise and coincidence. *Computational Linguistics*, (19):61–74.

Durbin, R., Szeliski, R., and Yuille, A. (1989). An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 1(3):348–358.

Edelman, S. and Weiss, Y. (1995). Hyperacuity. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 1009–1011. MIT Press.

Elman, J. (1991). Distributed representations, simple recurrent networks. *Machine Learning*, 7:195–225.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, (14):179–211.

Elman, J. L. (1993). The importance of starting small. *Cognition*, (48):71–99.

Erwin, E., Obermayer, K., and Schulten, K. (1992). Self-organizing maps: Ordering, convergence properties and energy functions. *Biological Cybernetics*, 67:47–55.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models.* Chapman and Hall.

Eysenck, M. W. and Keane, M. T. (1995). *Cognitive Psychology.* Lawrence Erlbaum Associates, 3rd edition.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications I.* Wiley.

Finch, S. (1993). *Finding Structure in Language.* PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Finch, S. and Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*

Firth, J. R. (1968). A synopsis of linguistic theory. In Palmer, F. R., editor, *Selected Papers of J. R. Firth: 1952-1959.* Longman.

Frean, M. (1990). The upstart algorithm: A method for constructing and training feedforward neural networks. *Neural Computation*, 2:198–209.

Fritzke, B. (1994). Growing cell structures – a self-organizing network for supervised and unsupervised learning. *Neural Networks*, 7:1441–1460.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58.

Georgopoulos, A. P., Kettne, R. E., and Schwartz, A. B. (1988). Primate motor cortex and free arm movement to visual targets in three-dimensional space. *Journal of Neuroscience*, (8):2928–2937.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems.* Houghton Mifflin.

Gillund, G. and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91(1):1–67.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations.* Johns Hopkins University Press, Baltimore, second edition.

Goodhill, G. J., Bates, K. R., and Montague, P. R. (1997). Influences on the global structure of cortical maps. *Proceedings of the Royal Society, Series B*, 264:649–655.

Goodhill, G. J. and Cimponeriu, A. (2000). Analysis of the elastic net model applied to the formation of ocular dominance and orientation columns. *Network*, 11:153–168.

Goodhill, G. J. and Richards, L. J. (1999). Retinotectal maps: molecules, models, and misplaced data. *Trends in the Neurosciences*, 22:529–534.

Goodhill, G. J. and Sejnowski, T. J. (1997). A unifying objective function for topographic mappings. *Neural Computation*, 9(6):1291–1303.

Goodhill, G. J. and Willshaw, D. J. (1994). Elastic net model of ocular dominance: Overall stripe pattern and monocular deprivation. *Neural Computation*, 6:615–621.

Graepel, T., Burger, M., and Obermayer, K. (1997). Phase transitions in stochastic self-organising maps. *Physical Review E*, 56(4):3876–3890.

Graepel, T., Burger, M., and Obermayer, K. (1998). Self-organizing maps: Generalizations and new optimisation techniques. *Neurocomputing*, (20):173–190.

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Magazine*, 1(2):4–29.

Halmos, P. R. (1987). *Finite-dimensional Vector Space.* Springer Verlag.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Society*, 84:502–516.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Addison-Wesley.

Hinton, G. E. and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. In *Philosophical Transactions of the Royal Society B*, volume 358, pages 1177–1190.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Science*, number 79, pages 2554–2558.

Huckle, C. C. (1996). *Unsupervised categorization of word meanings using statistical and neural network methods.* PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Jackson, J. E. (1991). *A User's Guide to Principal Components.* John Wiley and Sons.

Joliffe, I. T. (1986). *Principal Component Analysis.* Springer Verlag.

Kandel, E., Jessel, T., and Schwartz, K. (1991). *Principles of Neural Science.* Appleton Lange.

Karmiloff-Smith, A. (1995). *Beyond Modularity.* MIT Press.

Kaski, S. (1989). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of the International Joint Conference on Neural Networks*, pages 413–418.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM: Self-organizing maps of document collections. *Neurocomputing*, 21:101–117.

Keller, F. (1997). Extraction, gradedness, and optimality. In Dimitriadis, A., Siegel, L., Surek-Clark, C., and Williams, A., editors, *Proceedings of the 21st Annual Penn Linguistics Colloquium*, number 4.2 in Penn Working Papers in Linguistics, pages 169–186. Department of Linguistics, University of Pennsylvania.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian Inference.* Cambridge University Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, (43):59–69.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks*, 6:895–905.

Kohonen, T. (1995). *Self-organizing maps.* Springer, Berlin.

Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V., and Saarela, A. (1999). Self organization of a massive text document collection. In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 171–182. Elsevier.

Krekelberg, B. and Taylor, J. G. (1997). Nitric oxide: What can it compute? *Network: Computation in Neural Systems*, (8):1–16.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, (104):211–240.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842 1845.

Lin, X., Soergei, D., and Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In Bookstein, A., Chiaramella, Y., Salton, G., and Raghavan, V. V., editors, *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 262–269.

Livesay, K. and Burgess, C. (1998a). Mediated priming does not rely on weak semantic relatedness or local co-occurrence. In *Proceedings of the Cognitive Science Society*, pages 609–614.

Livesay, K. and Burgess, C. (1998b). Mediated priming in high-dimensional meaning space: What is mediated in mediated priming? In *Proceedings of the Cognitive Science Society*, pages 436–441.

Lorch, R. F. (1992). Priming and search processes in semantic memory: A test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior*, (21):468–492.

Lowe, W. (1997a). Meaning and the mental lexicon. In *Proceedings of the 15th International Joint Conference on Artificial Intelligenc*, pages 1092–1097. Morgan Kaufman.

Lowe, W. (1997b). Semantic representation and priming in a self-organizing lexicon. In Bullinaria, J. A., Glasspool, D. W., and Houghton, G., editors, *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 227–239, London. Springer-Verlag.

Lowe, W. and Blumstein, S. (2000). Modeling lexical access deficits in aphasia with a topographic lexicon. Unpublished manuscript.

Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665. Mahwah, NJ: Lawrence Erlbaum Associates.

Luttrell, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation*, 6(5):767–794.

MacKay, D. J. C. (1991). *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology.

MacKay, D. J. C. (1998). Gaussian processes: A replacement for supervised neural networks? Lecture notes, available from `http://wol.ra.phy.cam.ac.uk/mackay/`.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, (10):1–27.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Masson, M. E. (1991). A distributed memory model of context effects in word identification. In Besner, D. and Humphreys, G. W., editors, *Basic Processes in Reading: Visual Word Recognition*. Lawrence Erlbaum Associates.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (21):3–23.

McClelland, J. L. and Rumelhart, D. E., editors (1988). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press.

McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.

McKoon, G. and Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (18):1155–1172.

McKoon, G. and Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, (49):25–42.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*. MIT Press.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, (59):344–366.

Miikkulainen, R. and Dyer, M. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, (15):343–399.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294.

Morton, J. (1979). Word recognition. In Morton, J. and Marshall, J. C., editors, *Psycholinguistics Series 2: Structures and Processes*. Elek.

Moss, H. E., Hare, M. L., Day, P., and Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, (6):413–427.

Moss, H. E., Ostrin, R. K., Tyler, L. K., and Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (21):863–883.

Mulier, F. and Cherkassky, V. (1995). Self-organization as an iterative kernel smoothing process. *Neural Computation*, 7(6):1165–1177.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics No. 118, Springer-Verlag.

Neal, R. M. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. and Humphreys, G. W., editors, *Basic processes in reading: Visual word recognition*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68.

Oram, M. W., Földiák, P., Perrett, D. I., and Sengpiel, F. (1998). The ideal homunculus: decoding neural population signals. *Trends in the Neurosciences*, 21(6):259–265.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. In *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, pages 559–572.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 37–42. Mahwah, NJ: Lawrence Erlbaum Associates.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. E. (1994). Understanding normal and impaired word-reading: Computational principles in quasi-regular domains. Technical report, Carnegie Mellon University.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88:93–134.

Radford, A. (1988). *Transformational Grammar: a First Course*. Cambridge University Press.

Rao, R. P. N. and Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763.

Rasmussen, C. E. (1996). *Evaluation of Gaussian Processes and other methods for regression*. PhD thesis, Department of Computer Science, University of Toronto.

Ratcliff, R. and McKoon, G. (1981). Does activation really spread? *Psychological Review*, (88):454–462.

Ratcliff, R. and McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, (95):385–408.

Redington, M. and Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in the Cognitive Sciences*, 1(7).

Redington, M. and Chater, N. (1998). Connectionist and statistical approaches to language acquisition. *Language and Cognitive Processes*, 13.

Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, (61):241–254.

Ritter, H., Martinez, T., and Schulten, K. (1991). *Neural Computation and Self-Organizing Maps*. Addison Wesley, Reading MA.

Rojas, R. (1996). *Neural Networks—A Systematic Introduction*. Springer-Verlag, Berlin-New York.

Rose, K., Gurewitz, E., and Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594.

Roweis, S. (1998). EM algorithms for PCA and SPCA. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation*, 11(2):305–345.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.

Sammon Jr., J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, (18):401–409.

Schieber, S. M. (1986). *An Introduction to Unification-Based Approaches to Grammar*. CSLI.

Scholtes, J. C. (1993). Using extended feature maps in a language acquisition model. In *Proceedings of the Second Australian Conference on Neural Networks*, pages 38–43.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theory.* CSLI.

Shelton, J. R. and Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory and Cognition*, (18):1191–1210.

Shepherd, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, (27):219–246.

Spence, D. P. and Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, (19):317–330.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models.* PhD thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley.

Svensén, M. (1998). *GTM: The Generative Topographic Mapping.* PhD thesis, Neural Computing Research Group, Aston University.

Tipping, M. E. and Bishop, C. M. (1997). Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University.

Tonkes, B. and Wiles, J. (2000). Minimally biased learners and the emergence of compositional language. to appear in Proceedings of the 3rd Conference on the Evolution of Language.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, (17):401–419.

Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282.

Weiss, Y., Edelman, S., and Fahle, M. (1995). Models of perceptual learning in vernier hyperacuity. *Neural Computation*, (5):695–718.

Williams, C. K. I. (1998). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, K., editor, *Learning and Inference in Graphical Models.* Kluwer Academic Press.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo., M. E., editors, *Advances in Neural Information Processing Systems 8.* MIT Press.

Wilson, H. R. (1986). Resposes of spatial mechanisms can explain hyperacuity. *Vision Research*, (26):453–469.

Wittgenstein, L. (1958). *Philosophical Investigations.* Blackwell.

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–430.

Zipf, G. K. (1949). *Human Behavior and the Principal of Least Effort.* Addison Wesley.

Appendix A

Basis Elements

ability academic accept access account activity add addition admit advantage advice affair age agency ago agree aid alternative amount animal annual answer apparently appeal apply area army arrange article associate association attack attempt attention average avoid aware balance bar base battle bed begin benefit big bill black block blood board body book box branch bring broad brother build building business buy carefully carry case catch central century challenge chance change chapter character charge check child choice choose church city class clean clear close club college comment commercial committee common communication compare complete completely complex concentrate condition confirm contact content continue contract contrast cost country county couple cover create cross customer cut damage date day deal death debate decide deep defence demand depend describe develop direct direction discover display distance district doctor door double doubt draw drive drug due duty early earth easily east easy economy education emerge employ encourage end enter equally equipment essential evening event eventually evidence exercise exist expect experience explain extend extent extra extremely facility fact fail failure fair fall family favour fear feature feeling female field fight figure fill film finally financial find fit flow follow force foreign forward free friend front full fully future general give glass god good government grant green ground growing growth half handle happy health hear heart high history hit hold hope hospital hotel house huge husband idea identify ignore image immediately improve include including industrial industry influence intend interest involve island job join judge key kind king knowledge land language large largely law lead leader learn left letter lie life line list listen living lord low magazine maintain major make male man manage management manner mark marriage mass master material matter measure medical meet meeting member memory mention message method middle mile military mind minister minute modern moment money mother move movement music natural nature news newspaper normal north northern note notice number observe occasion occur offer office officer official open operate opportunity original page paper parent park part pass past pattern pay people period person personal physical pick picture piece plan planning plant point police policy political poor popular population position positive post power powerful practical practice prepare presence present press pressure prevent problem produce professional programme promise property proportion

protect prove provided public pull purpose question quickly raise range reach read reader reading ready real reality reason receive recent recently recognise reduce reflect refuse regard regular relate relationship relative remove replace reply represent research respect responsibility result retain return reveal rich ring rise risk role room royal run safe scale scene scheme school science sea search seat secretary section sector secure security seek sell send sense serve service sex shape share shop short show sight sign similar simple simply single site situation size skill slightly slow small south space speak special spend spread spring staff stand standard star station status step stone story street strength strike strong student study style subject suggest survey table target task tax technique telephone tend term test theory thought time today touch town traditional training transport travel treat treatment true trust turn type understand university usual variety village visit wait wall watch water wave week west white woman wood word work worker working works world worth write writer year

Appendix B

Notation

Numbers

All numerical quantities are integer or real-valued. The set of real numbers is denoted $\mathcal{R}$.

Lower case roman letters denote scalars. The letters x, y, t and b are variables. All others are constants. Upper case roman letters are constant scalars, usually denoting dimensionality e.g. L-dimensional space, or the maximum values of sums.

Vectors are denoted by bold lower case greek or roman letters e.g. $\boldsymbol{\mu}$. All vectors are column vectors unless defined otherwise. Matrices are denoted by bold upper case greek or roman letters e.g. $\mathbf{H}$. The i,jth element of $\mathbf{H}$ is denoted $\mathbf{H}_{ij}$. the transpose of $\mathbf{H}$, $\mathbf{G} = \mathbf{H}^{\mathsf{T}}$ has $\mathbf{G}_{ij} = \mathbf{H}_{ji}$. $\mathbf{I}$ is the identity matrix, a square matrix containing zeros except for the main diagonal elements which take the value one. containing only zeros. The inverse of $\mathbf{H}$ is denoted $\mathbf{H}^{-1}$, where

$$\mathbf{H}\mathbf{H}^{-1} = \mathbf{H}^{-1}\mathbf{H} = \mathbf{I}$$

If not stated in the text the exact dimensions of a vector or matrix are implied by the algebraic context in which it is used.

Functions

A function taking a scalar argument x is denoted $f(x)$, or occasionally by a greek letter instead of f.

In functional contexts D is an operator: $D^i f(x)$ denotes the ith derivative of the function f with respect to x. $f'(x)$ denotes the first derivative of f with respect to x. More primes mark higher derivatives.

Functions taking vector or matrix arguments are denoted by upper case roman or greek letters. Whether the range of the function is scalar or multivariate is determined from algebraic context.

δ_{ij} denotes Kronecker's delta, a function that takes the value 1 when $i = j$ and 0 otherwise.

Probability

Although a random variable X is realised by particular values x we denote both the distribution function mapping possible values onto [0,1] and the probability of a particular value, x, as $p(x)$. Whether $p(x)$ denotes a probability or a probability distribution is determined by context. Likewise $p(x, y)$ denotes either the probability that simultaneously $X = x$ and $Y = y$, or the joint distribution of X and Y. $p(x \mid y)$ denotes either the probability that $X = x$ given that $Y = y$ or the the distribution of X given that $Y = y$.

Calligraphic letters are used to specify function forms for probability distributions. If X is a random variable in $\mathcal{R}^d$ then

$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{C})$$

means that $\mathbf{X}$, is distributed according to a Normal distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{C}$. Then

$$p(\mathbf{x}) = (2\pi)^{-d/2} \mid C \mid^{-1/2} \ \exp(-1/2\ (\mathbf{x} - \mathbf{a})^{\mathsf{T}}\ \mathbf{C}^{-1}\ (\mathbf{x} - \mathbf{a})).$$

If X is a random variable with two possible values e.g. $x \in \{1,0\}$ where $p(X = 1) = \theta$, then $p(X = 0) = 1-\theta$ and the distribution of possible values in a sample of size N is Binomial with parameters θ and N,

$$x \sim \mathcal{B}(\theta, N)$$

and

$$p(x) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}.$$

Binomial distributions are useful as models of occurrence probabilities when e.g. $x \in$ {'word w occurs', 'word w does not occur'}.

The uniform distribution between points a and b in $\mathcal{R}^N$ is denoted $\mathcal{U}(a, b)$. For $N > 1$ we adopt the convention that the range of $\mathcal{U}$ is the set of points in *each* dimension greater than a and less than b. For example $\mathcal{U}(-1, 1)$ is a square centred on the origin in $\mathcal{R}^2$, and a cube centred on the origin in $\mathcal{R}^3$.